PREDICTIVE SUMMARISATION IN EVO 2

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic causal language models learn representations through next-nucleotide prediction, yet these representations are also used to infer biological properties of DNA. We investigate how biological readout and predictive function diverge across depth in Evo 2 7B and 40B. Both models exhibit an abrupt representational transition at which variation across genomic positions concentrates in fewer directions. Across this transition, the decline in genomic annotation readout steepens while next-base probe performance improves, including when both are evaluated at the same positions. Restricting standardised transition states to their 16 leading principal directions and continuing pretrained computation preserves native next-base prediction within prespecified NLL and output-KL tolerances, while reducing annotation AUROC by 0.086 in 7B and 0.059 in 40B relative to intact transition states. The same restriction before the transition fails to preserve prediction. Annotation decline also persists across probe controls and among positions with similar predictive distributions. We interpret these findings as *predictive summarisation*: a reorganisation of contextual representations for next-base prediction that reduces the accessibility of genomic annotations. Earlier representations selected during development outperform final embeddings by 0.115 in mean annotation AUROC in a retrospective chromosome 19 evaluation, with benefits extending to gene-structure and conservation readouts. Some variant-difference readouts instead favour final embeddings, showing that layer preference depends on the biological target and feature construction. Next-nucleotide prediction provides indirect supervision for biological properties; representations organised for this objective need not be the most useful for biological analysis.

1 INTRODUCTION

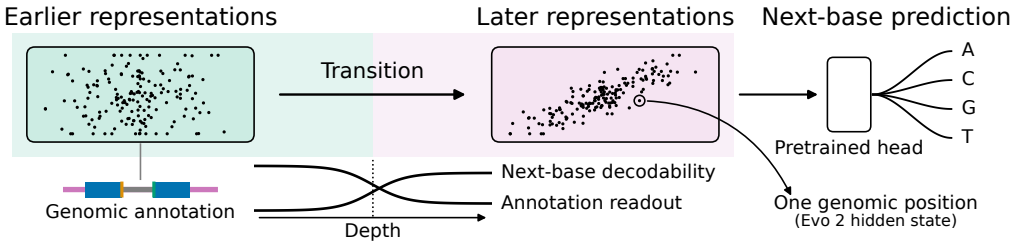

Figure 1: Predictive summarisation in Evo 2 (schematic). Across the transition (b28 in 7B, b21 in 40B; zero-indexed blocks), variation across genomic positions concentrates while annotation readout decreases and next-base decodability improves. The pretrained head reads the final representation; the inset shows schematic depth trends. Dot clouds are illustrative; measured trajectories follow in Figures 3 and 5.

Genomic causal language models learn contextual representations through DNA sequence prediction, supporting analyses of genomic regions, evolutionary constraint and variant effects (Brixi et al., 2026). This supervision is indirect: next-base prediction does not explicitly require distinguishing protein-coding exons from introns removed by splicing. Biological features learned this way therefore need not remain accessible to fitted decoders throughout the model's depth.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Genomic studies find depth-dependent representation quality (Dalla-Torre et al., 2025; Tang et al., 2025), and language models show depth-dependent embeddings (Skean et al., 2025), massive activations (Sun et al., 2024) and dominant spectral directions (Wang et al., 2025; Queipo-de Llano et al., 2026). In Evo 2, Hirota et al. (2026) evaluate layer-dependent readouts on biosynthetic gene clusters and include logit-lens analyses, and Arc Institute (2026b) reports high-norm later layers with little additional predictive benefit, deriving Evo 2 20B from 40B by model surgery. Complementing these analyses, we ask how annotation readout and the model's own next-base prediction respond to the same restriction of a hidden state.

We find an abrupt transition in both Evo 2 7B (b28) and 40B (b21): variance across positions concentrates, annotation decline steepens relative to preceding blocks, and next-base decodability improves across the transition. We interpret this joint pattern (Figure 1) as *predictive summarisation*: reorganisation of contextual representations for next-base prediction, with reduced annotation accessibility. These changes precede, in depth, the strongest measured alignment with the output head.

Keeping only 16 leading principal directions of the transition state preserves the model's own next-base prediction within prespecified tolerances but lowers annotation AUROC by 0.086 (7B) and 0.059 (40B); the same restriction before the transition fails to preserve prediction. Annotation decline after the transition persists among positions with similar predictive distributions, and tested attenuations of transition-block updates do not restore annotation without harming prediction. Earlier representations outperform final embeddings on annotation and conservation readouts, whereas some variant-difference readouts favour final embeddings. These findings motivate selecting extraction depth for the intended biological target and feature construction.

2 Models and Measurements

2.1 Per-position states and functional readouts

Frozen models process 6 kb windows of the human reference genome GRCh38 (hg38) (Schneider et al., 2017) in increasing reference-coordinate order. Each causal state uses only sequence through its position: annotation labels describe that position; next-base targets describe the following, unseen base. Evo 2 7B/40B have 32/50 StripedHyena 2 blocks of widths 4096/8192 (Brixi et al., 2026; Ku et al., 2025). We prefix zero-indexed block outputs with b; `norm` denotes the separate final-RMSNorm output, our *final embedding*. States remain per-position vectors without pooling. The main 40B trajectories use FP8 (Appendix E.3).

Fitted probes read annotations or next bases from each state; native prediction continues the remaining pretrained blocks and original head (Table 1). Unedited continuation gives the same final output from any depth. Interventions can change predictive performance (NLL) and output distributions (KL).

Table 1: Three questions about a state. Annotation probes report AUROC; next-base probes, accuracy and NLL. Native evaluation uses next-base NLL and intact-to-edited four-base KL.

Measurement	Question	Evaluation
Annotation readout	How well can genomic annotations be decoded?	Separately fitted annotation probe
Next-base decodability	How well can the next base be decoded?	Separately fitted next-base probe
Native prediction	What does the pretrained model predict from this state?	Remaining blocks and original head

Coding exons contain protein-coding sequence. Splicing removes intervening introns and joins exons; donor and acceptor sites mark intron starts and ends in transcript orientation (Figure 2). GENCODE v44 gene annotations (Frankish et al., 2023) supply splice labels covering the two intronic bases at each boundary.

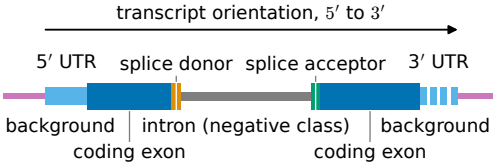

Figure 2: Gene structure (schematic).

Genes occur on both strands, so donor/acceptor order reverses for reverse-strand genes in our input order. Four binary tasks each contrast one class—coding exon, donor, acceptor or background—with intron positions, excluding positions assigned another label from the intron class. Background means positions without a retained label and can include noncoding exonic bases. The cross-scale summary averages these four AUROCs. Some 7B evaluations also include $5'/3'$ untranslated regions (UTRs), transcript regions outside the coding sequence.

Unless specified otherwise, each state separately fits training-fold coordinate standardisation, PCA to 256 components (PCA256) and logistic regression. Five-fold cross-validation groups windows connected by any chain of genomic overlaps. The named annotation class is positive; held-out scores retain their training-defined orientation.

Within each assay, sampled genomic positions, their input windows and folds stay fixed across states. The primary annotation and next-base assays use different samples; a separate comparison shares positions and training-fold preprocessing between the two probes within each model (Appendix B.5).

Unless stated otherwise, brackets give paired 95% intervals that reweight groups of overlapping input windows, with fitted probes and extraction choices held fixed. They describe variation across the evaluated groups, excluding probe-refitting and layer-selection uncertainty.

Original chr22/chr17 panels support development and extraction-state selection. Chr19 is retrospective because earlier outcomes were inspected during development. A 7B chr18 panel was fixed before its readout outcomes, although chr18 had appeared in earlier analyses. Extraction choices transfer across panels; probes are refitted within each (Table A3).

2.2 Geometry and transition locations

Geometry uses the same 2048 randomly sampled chr22 positions across states within each model. The *adjacent norm ratio* compares mean vector norms with the preceding block. *Standardised effective rank* (Roy & Vetterli, 2007) measures variance concentration after centring and coordinate standardisation: it equals the number of directions when they share variance equally and approaches one as a single direction dominates. *Output-head alignment* measures the fraction of centred, unstandardised variance in the three independent A/C/G/T logit-contrast directions: four logits have three independent contrasts after removing their common offset. Rank equalises coordinate scales; alignment retains them (Appendix A.5).

Original profiles show sharp norm increases and variance concentration at b28 (7B) and b21 (40B) (Figures 3 and 5). These observed *transitions* remain fixed throughout our comparisons; minimum standardised effective rank occurs later, at b30/b23.

The b21 expansion also appears on a uniformly spaced 40B position panel ($\rho = 235$; Appendix E.3).

3 Predictive summarisation in Evo 2

3.1 Opposing readout changes at the transition

We compare the fitted annotation and next-base probes across b27→b28 (7B) and b20→b21 (40B). At these transitions, mean state norms rise over 200-fold and standardised effective rank falls over an order of magnitude (Figure 3(a,b)). Annotation readout declines while next-base decodability improves (Figure 3(c,d)).

Four-task annotation AUROC changes by -0.038 $[-0.048, -0.025]$ in 7B and -0.036 $[-0.047, -0.027]$ in 40B. Next-base accuracy instead rises by 0.019 $[0.009, 0.028]$ and 0.027 $[0.019, 0.035]$; probe NLL also improves.

Subtracting the mean of the preceding three adjacent changes gives the *excess change*: -0.036 $[-0.047, -0.022]$ in 7B and -0.027 $[-0.038, -0.017]$ in 40B for annotation. Thus annotation decline steepens (Figure 3(e)). Accuracy increases across the transition without a resolved acceleration: both excess-change intervals include zero (Figure 3(f)). The transition marks local steepening, rather than the largest annotation loss: the later b28→b29 and b22→b23 drops are larger by point estimate in 7B and 40B (Appendix B.2). Transition-to-final AUROC falls by a further $0.079/0.072$.

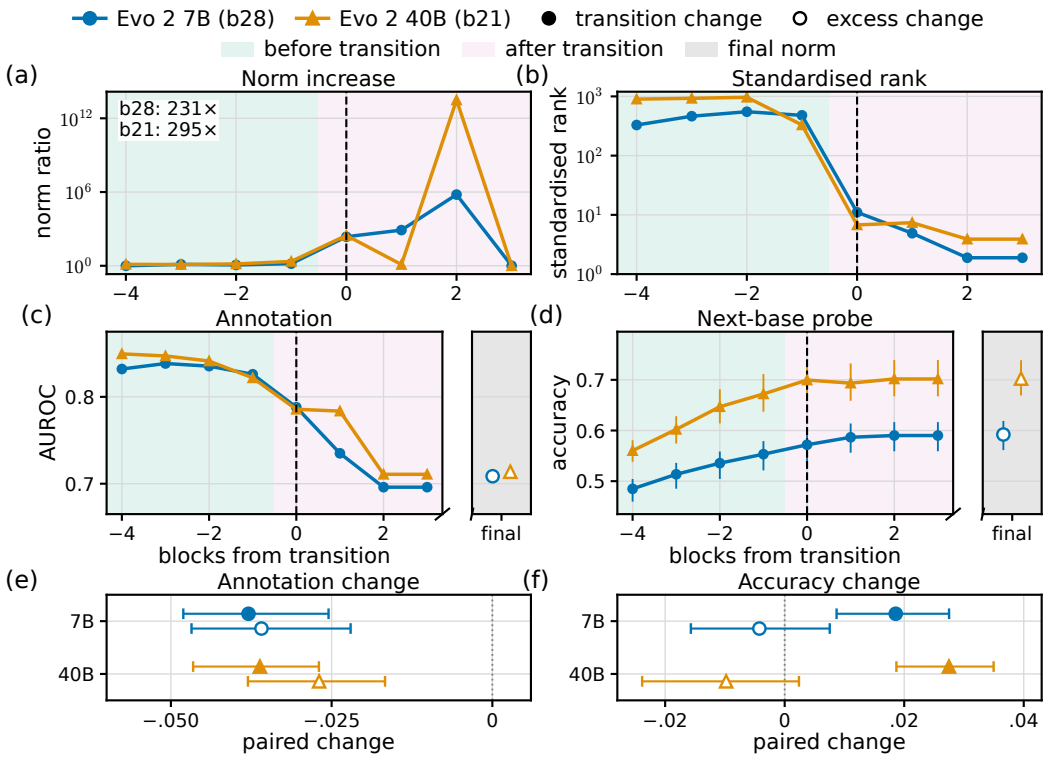

Figure 3: Geometric and functional changes around b28 (7B) and b21 (40B). (a,b) Chr22 adjacent norm ratios and standardised effective rank. (c,d) Four-task mean annotation AUROC and next-base probe accuracy. Dashed lines mark transitions; tints distinguish earlier/later states and final normalisation (open markers). Breaks omit intervening blocks, including 40B b25–b49. (e,f) Transition changes (filled) and changes relative to the preceding three-change mean (open, *excess change*). Bars are conditional paired 95% genomic-group intervals; only next-base trajectories have per-state intervals. Assays use different positions, fixed across states.

Both probes on identical positions, with shared standardisation and PCA, retain the opposing changes at 6 kb (annotation $-0.030/-0.037$, accuracy $+0.011/+0.021$ in 7B/40B, with accuracy gains below the preceding-change reference) and at 24 kb in 7B (Appendix B.5).

Ten probe protocols retain negative transition and pre-to-final intervals, including without PCA and with training-selected regularisation (Table 2); local steepening uses the primary protocol. On the 32-window dimension-restriction panel, validation-stopped MLPs reach the stopping criterion in all 80 fits per model and retain the transition decreases ($0.040/0.042$ in 7B/40B). Relative to linear probes, the subsequent transition-to-minimum-rank decline is smaller by $0.030/0.042$, but pre-to-final gaps of $0.082/0.097$ remain. Secondary post-PCA scaling reverses the initial contrast (Appendix B.4).

Table 2: Annotation decline persists across probe controls. Changes in four-task mean AUROC across the transition (transition state minus immediately preceding state), with paired 95% intervals. The primary, no-PCA and training-selected protocols are shown; Table A9 reports all ten.

Probe protocol	*Evo 2 7B*	*Evo 2 40B*
Primary PCA256	-0.038 $[-0.048, -0.025]$	-0.036 $[-0.047, -0.027]$
No PCA, fixed regularisation	-0.023 $[-0.033, -0.011]$	-0.042 $[-0.056, -0.029]$
No PCA, training-selected regularisation	-0.021 $[-0.028, -0.012]$	-0.031 $[-0.041, -0.021]$

The 24 kb context-length assay prepends 18 kb to the same 7B windows, keeping readout endpoints fixed but sampling next-base positions separately from the shared-position assay. Concentration,

annotation steepening and NLL improvement persist, but its accuracy change is unresolved (Appendix B.6).

A second, disjoint 32-window chr22 panel, with inputs and size fixed before model evaluation, replicates the annotation decline: pre-transition → transition → minimum-rank AUROC is $0.842 \rightarrow 0.807 \rightarrow 0.697$ in 7B and $0.853 \rightarrow 0.795 \rightarrow 0.708$ in 40B (Appendix C).

On both 32-window panels, transition-to-minimum-rank and pre-to-final declines persist when positions are stratified by predicted nucleotide and entropy or matched by four-base predictive distributions. On the dimension-restriction panel, the initial 40B decrease is resolved by pooled stratification and matching, but not by unstratified readout or within-fold stratification; on the second panel it is resolved throughout (Appendix C).

3.2 Prediction and annotation differ under dimension restriction

Does retaining only a state's leading principal directions preserve native prediction and annotation readout together? We project each position's coordinate-standardised state onto its leading k directions, restore scales and mean, and continue pretrained computation. The edit removes variation outside these directions while preserving hidden width (Figure 4).

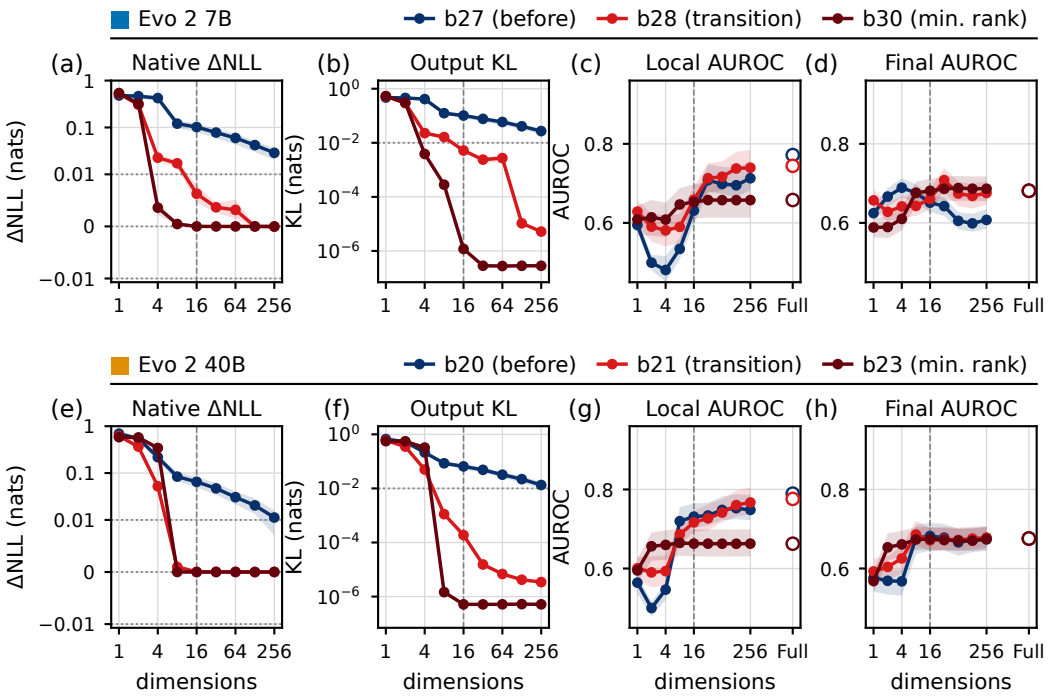

Figure 4: At the transition, 16 leading standardised principal directions retain prediction while reducing annotation readout. (a–d) 7B b27/b28/b30; (e–h) 40B b20/b21/b23. Panels (a,e) show native NLL changes, (b,f) output KL, (c,g) local AUROC and (d,h) final AUROC. Native NLL changes (edited minus intact), intact-to-edited KL and refitted annotation probes use identical positions in 32 disjoint windows. Local probes read restricted states; final probes read their continued final embeddings. Scores average folds and four tasks equally. *Full* marks unedited states. Bands are pointwise 95% genomic-group intervals conditional on fitted probes and bases, paired for changes. Native retention requires the NLL-change interval within ± 0.01 nats and the KL upper bound at most 0.01 nats (dotted guides). NLL uses a symmetric log axis around ± 0.01; log KL floors values below 10^{-7}. Vertical dashed guides mark $k = 16$.

Means, scales and PCA bases are fitted on 12 basis-fitting windows. The 32 windows of the dimension-restriction panel are mutually disjoint and overlap neither these inputs nor the original chr22 panel. Native NLL, intact-to-edited output KL and annotation probes use identical positions within each

model. Probes are refitted on the restricted state (*local AUROC*) and on the final embedding produced by continuing from it (*final AUROC*) (Appendix D).

Each edited state is compared with its own intact baseline. Unlike the primary panel, the intact 40B transition contrast here is unresolved (-0.0133 $[-0.0346, 0.0041]$ AUROC). At $k = 16$, 7B b28 changes NLL by $+0.0063$ $[0.0045, 0.0079]$ nats, with KL 0.0052 $[0.0044, 0.0058]$. Both meet the prespecified margins: the NLL-change interval lies within ± 0.01 nats and the KL upper bound below 0.01 nats. Local annotation AUROC falls from 0.744 to 0.659 (-0.086 $[-0.122, -0.065]$). At 40B b21, 16 directions also retain prediction (NLL change $+0.00002$ nats; KL 0.00019), while local AUROC falls from 0.776 to 0.717 (-0.059 $[-0.083, -0.044]$). The restriction therefore retains native prediction within the margins (Figure 4(a,b,e,f)) while reducing local annotation accessibility (Figure 4(c,g)).

Before the transition, the same restriction costs $0.102/0.065$ nats in 7B/40B and fails native retention. At the later minimum-rank states b30/b23, 16 directions retain both native prediction and local annotation. Four directions still cost 0.343 nats at 40B b23 despite its intact effective rank of 3.75 on this panel: effective rank does not specify the dimension needed for native function (Figure 4(e)).

3.3 Rank decreases before the strongest output alignment

Both models show variance concentration at the transition and later minimum standardised effective rank (7B b30, 1.89; 40B b23, 3.89; Figure 5(b)), followed by maximum output-head alignment after final RMSNorm (Figure 5(c)). In 40B, major changes cluster at b21–b23; subsequent stored states change little except b34 and final RMSNorm. At 7B b31, finite-precision addition absorbs late branch writes into the larger residual (Appendices E.1 and E.3).

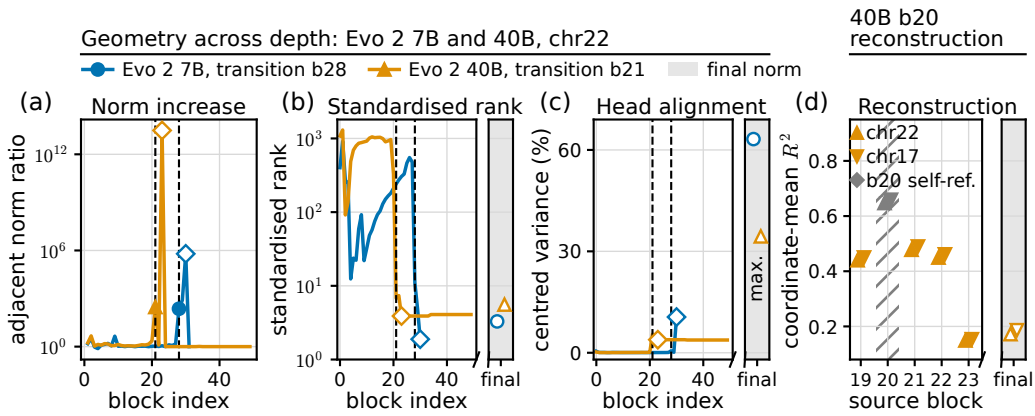

Figure 5: Staged geometry (a–c) and earlier-state reconstruction (d). (a–c) Adjacent norm ratio, standardised effective rank and centred ACGT-contrast alignment for Evo 2 7B and 40B on chr22; dashed lines and filled markers mark the transitions b28/b21, open diamonds the minimum-rank blocks b30/b23, and the tinted `final` column the post-RMSNorm embedding, each model's largest measured alignment. (d) Coordinate-mean R^2 for reconstructing 40B b20 from each source with training-fold PCA256 and ridge ($\alpha = 1$), averaged over five folds on chr22 (▲) and chr17 (▼); no intervals are retained for this trajectory. Intervals for separately tuned decoders appear in Table A33. Grey b20 points are the self-reference, the same protocol predicting held-out b20 coordinates from b20, a calibration rather than an information fraction.

Intermediate alignment trajectories differ. In 40B it decreases from 5.47% at b21 to 3.86% at b23, before reaching 34.43% at final normalisation; 7B reaches its maximum of 63.21% there.

A comparison under low rank. For CENO 1B, a separately trained genomic language model (Ma et al., 2026; CladeTeam, 2026), we measure a final-block annotation AUROC increase from 0.653 to 0.781 while standardised effective rank rises from 2.19 to 5.13. This recovery precedes the larger rank increase at final normalisation (Appendix F.1). Low rank alone therefore does not imply reduced annotation accessibility; the Evo 2 interpretation rests on the joint functional pattern.

3.4 EARLIER-STATE STRUCTURE REMAINS PARTLY RECOVERABLE

To test whether earlier structure survives concentration, we reconstruct the 8192 coordinates of the 40B pre-transition state b20 using training-fold PCA256 and ridge, assessing coordinate-mean R^2 over five overlap-connected folds (Figure 5(d); Appendix G).

The b20 self-reference uses the same 256-component bottleneck and held-out prediction, so R^2 need not equal one. Against this reference ($0.647/0.659$ on chr22/chr17), earlier-state structure remains partly recoverable at b21 ($0.478/0.490$) and b22 ($0.450/0.461$). Reconstruction drops sharply at b23 ($0.149/0.154$) and rebounds slightly at the final embedding ($0.172/0.186$), without returning to its post-transition level.

The b21–b23 gap persists with PCA64, PCA256, PCA1024 and no projection, using training-selected ridge penalties: every paired interval is below zero on both chromosomes (Table 3). A validation-stopped nonlinear decoder retains the chr22 decline ($R^2 = 0.602$ to 0.139; Appendix G). The later loss of recoverability therefore persists across decoder families and bottlenecks.

Table 3: Reconstruction of 40B b20 declines from b21 to b23 across source dimensions with training-selected ridge penalties. Entries are mean coordinate-wise R^2; all eight paired b23-minus-b21 95% intervals lie below zero (Table A33).

	chr22		*chr17*	
Source PCA	b21	b23	b21	b23
64	0.333	0.149	0.341	0.154
256	0.478	0.149	0.490	0.154
1024	0.644	0.149	0.652	0.154
No PCA	0.593	0.149	0.645	0.154

3.5 NATURAL-UPDATE ATTENUATION TESTS ANNOTATION RECOVERY

State restriction tests which variation downstream prediction can forgo. We next ask whether weakening a natural transition update can recover annotation readout while retaining prediction. We multiply either the mixer or gated-MLP write by α before residual addition and recompute the remaining model. Reusing the dimension-restriction panel for dose selection, we select the smallest $\alpha < 1$ that meets native-retention margins at itself and all larger tested attenuated values. We then evaluate these doses on the second, disjoint 32-window panel (Appendix H.1).

The selected MLP/mixer doses are $0.02/0.35$ in 7B and $0.85/0.35$ in 40B. All four retain native prediction, but none meets the final-annotation recovery criterion (a gain of at least 0.01 AUROC with a positive paired interval and Holm-adjusted $p \leq 0.05$). Final AUROC changes are $+0.0023/-0.0086$ in 7B and $-0.0004/+0.0038$ in 40B; the 7B mixer decrease has an interval below zero, so retention does not imply annotation invariance (Table A34).

Across the tested transition-write attenuations, no condition meets both recovery and native-retention criteria. Larger gains occur at prediction-damaging doses: in 40B, MLP scaling by 0.01 raises final AUROC by 0.056 $[0.040, 0.080]$ while increasing native NLL by 0.330 $[0.290, 0.374]$ nats (Figure A6(c,d)). These dose-sweep comparisons are descriptive. Removing earlier 40B control branches also improves annotation while worsening prediction, so this trade-off is not unique to the transition (Appendix H.1).

Branch ablations and component edits provide complementary evidence: transition and later branches contribute to native prediction, and retaining only the leading raw-update component fails to retain it (Appendices H.2 and H.4). Together with state restriction, the attenuation results distinguish removing annotation-relevant variation from recovering it by weakening the model's own updates.

4 DEPTH-DEPENDENT BIOLOGICAL READOUT

4.1 EARLIER STATES SUPPORT ANNOTATION AND CONSERVATION

We evaluate extraction states b24 (7B) and b17 (40B), chosen using chr22/chr17 annotation readouts to calibrate a shared four-block offset before the fixed transitions. These serve downstream readout; b27/b20 instead anchor local transition contrasts. Extraction states stay fixed while probes are refitted within each evaluation panel (Appendix I.4).

On chr19, earlier states improve mean annotation AUROC by 0.115 $[0.100, 0.133]$ over final embeddings (Table 4). The mean equally weights ten model–task cells: six at 7B, including UTRs, and four at 40B; the other rows of the table average the listed tasks within one model. Chr19 is retrospective. On the separately fixed 7B chr18 panel, the gain is 0.124 $[0.104, 0.142]$ (evaluation roles in Section 2.1; Appendix I.1).

Table 4: Earlier states (7B b24, 40B b17) improve sequence-property readout over final embeddings. The first four rows report mean annotation AUROC (7B adds both UTR tasks); BEND reports nine-class MCC. Differences use unrounded scores with paired 95% intervals. Chr19 is retrospective; the chr18 panel was fixed in advance.

Evaluation	Tasks/classes	Earlier	Final	Earlier − Final
■ ■ chr19 (7B/40B)	$6 + 4$	0.856	0.740	$+0.115$ $[+0.100, +0.133]$
■ chr19 (7B)	6	0.844	0.756	$+0.087$ $[+0.068, +0.109]$
■ chr19 (40B)	4	0.873	0.716	$+0.158$ $[+0.142, +0.176]$
■ chr18 (7B)	6	0.860	0.735	$+0.124$ $[+0.104, +0.142]$
■ BEND subset (7B)	9 classes	0.853	0.565	$+0.288$ $[+0.260, +0.317]$

We extend the binary readouts to multiclass gene structure using a filtered, class-subsampled BEND gene-finding subset (Marin et al., 2024). Here b24 improves nine-way region-and-strand classification by 0.288 $[0.260, 0.317]$ in Matthews correlation coefficient (MCC; Table 4). Its separate split precludes direct comparison with published benchmark scores.

Conservation tests a continuous property: phyloP100way scores evolutionary constraint across 100 vertebrates (Pollard et al., 2010). All twelve earlier-minus-final Spearman comparisons (two models, two chromosomes, three position subsets) are positive; eleven intervals exclude zero. Noncoding gains are 0.120–0.166 across scales and chromosomes (Figure A7, Appendix I.3).

The four-block offset is a development choice: recalibration with overlap-connected folds selects nearby b25/b18, which remain competitive (Appendix I.4). For a new target, compare earlier and final states on a validation panel and fix the state and feature choice before test evaluation.

4.2 Variant-difference readouts show different layer preferences

Reference-sequence features h describe a sequence at a position; variant-difference features $\Delta h = h_{\text{alt}} - h_{\text{ref}}$ compare a mutated sequence with its reference. We read Δh at the altered position or average it over all 6000 positions in the window. Earlier positions are unchanged by causality, so only positions at and after the variant contribute to this full-window average.

We compare 7B b24 and the final embedding with separately fitted probes. ClinVar's clinical variant annotations (Landrum et al., 2018) contrast pathogenic with benign missense variants (changes that alter a protein's amino acid); the GPN-MSA panel (Benegas et al., 2025) contrasts pathogenic with common missense variants. For BRCA1, probes predict experimentally measured variant effects from saturation editing (Findlay et al., 2018); higher scores indicate better preserved function. This evaluation contains five nonoverlapping input-context groups within one gene. Feature formulas, grouping and fitting procedures appear in Appendix J.1; Table 5 reports the resulting comparisons.

ClinVar pooled-difference features and both GPN-MSA feature configurations favour final embeddings. BRCA1 position-difference features favour b24. The intervals for ClinVar position differences and BRCA1 pooled differences include zero (Table 5). Layer preference varies with the target–feature combination, including among position-difference features. These comparisons concern the two fixed extraction states.

For context, the GPN-MSA predictor, which uses multispecies sequence alignments, scores higher on matched variants and folds (Appendix J.2). Within Evo 2, final-state advantages show that reduced annotation accessibility can coexist with useful variant-difference readouts.

Table 5: Variant-difference readouts prefer different states. Absolute scores compare 7B b24 with the final embedding; positive differences favour b24, and bold marks the state favoured by a paired 95% interval excluding zero. r_s is Spearman correlation.

Target (metric)	Feature	Earlier	Final	Earlier − Final
ClinVar (AUROC)	Position Δh	0.846	0.834	$+0.013$ $[-0.006, +0.033]$
ClinVar (AUROC)	Pooled Δh	0.885	**0.903**	-0.019 $[-0.033, -0.005]$
GPN-MSA panel (AUROC)	Position Δh	0.838	**0.866**	-0.028 $[-0.049, -0.002]$
GPN-MSA panel (AUROC)	Pooled Δh	0.902	**0.938**	-0.036 $[-0.054, -0.018]$
BRCA1 (r_s)	Position Δh	**0.285**	0.182	$+0.103$ $[+0.055, +0.200]$
BRCA1 (r_s)	Pooled Δh	0.390	0.359	$+0.032$ $[-0.031, +0.105]$

5 Discussion

Our findings distinguish learning biologically useful features from maintaining their accessibility during prediction. In Evo 2, annotation readout declines while next-base decodability improves across the transition; state restriction preserves native prediction within tested tolerances while reducing local annotation accessibility. We interpret this joint pattern as predictive summarisation.

Predictive summarisation as functional reorganisation. The restriction results show that annotation-supporting variation can be removed while native prediction remains within tested tolerances. This reorganisation unfolds in stages: concentration precedes the sharper loss of earlier-state recoverability in 40B and the strongest output-head alignment. CENO's annotation recovery under low rank shows that spectral concentration alone does not determine biological accessibility. Attenuating natural updates does not recover final annotation at prediction-retaining doses, leaving the causal relationship between the natural transition and annotation decline unresolved.

Biological utility depends on the readout. Earlier states support annotation and conservation, whereas some variant-difference readouts favour final embeddings. Reference features describe a sequence; difference features measure the model's response to a mutation. Later states may expose predictive-context changes despite reduced annotation accessibility, requiring joint measurement of hidden-state and native-prediction changes. BRCA1 position differences favour earlier states, so variant readouts also require target-specific layer selection. Extraction depth and feature construction should therefore be selected jointly for each biological target and fixed before evaluation.

From sequence statistics to biological function. Observed DNA reflects biological constraints, allowing sequence prediction to provide indirect biological supervision (Brixi et al., 2026). Yet its target is a nucleotide distribution, whereas biological studies may seek cell-specific activity or mutational consequences. The objective does not specify which biological quantities a downstream readout should recover. Our results locate this distinction within a trained model: distinctions decoded at one depth can become less accessible as prediction proceeds. Biological transfer therefore depends on both what sequence learning encodes and how subsequent computation presents it.

Biologically distinct contexts can imply similar next-base distributions, so prediction at one position need not keep their functional differences readable. This does not require low-rank hidden states: computing the distribution may require rich representations, and later predictions or mutational responses can still differ. The design question is which biological distinctions an objective rewards. Bidirectional context changes available sequence evidence (Schiff et al., 2024); functional supervision changes targets, as in AlphaGenome, within assayed cellular contexts (Avsec et al., 2026). Combining both warrants evaluation of predictive performance and biological accessibility.

Limitations and future directions. Evidence covers two frozen Evo 2 scales, mainly human 6 kb windows, and selected biological panels. Probe scores measure decoder-dependent accessibility, not total information. Chr19 is retrospective; intervals condition on fitted probes and extraction choices; BRCA1 covers five context groups. Training trajectories, selective interventions and broader assays could test how this separation arises and generalises, toward representations that support sequence prediction while keeping biologically required distinctions accessible.

AI USE STATEMENT

In this work, we used generative AI tools to write code, to provide feedback on the experiments and to assist in interpreting results. We have not used generative AI tools to generate synthetic data sets or to formulate or prove mathematical claims, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to run experiment code, to identify related literature, to draft parts of the paper and to edit it for readability. We have reviewed all AI-assisted work. Written protocols and amendments distinguish prospective choices from outcome-informed development steps. Reported statistics were recomputed by independent validation scripts from stored predictions, the central intervention was re-executed from the released code, and manuscript numbers were checked automatically against the stored results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Code, execution environments, experimental settings and data definitions are available at `https://anonymous.4open.science/r/ps-reproduction-public`. The repository contains the scripts that produced every reported result, their fixed configurations and recorded job commands, the panel coordinates with rows, labels, groups and folds, the software environments, checkpoint revisions and provider patches, and the reported summary values for comparison with a re-run. Genome sequence, model weights and activation caches are not redistributed; the panel sequences are regenerated from GRCh38 by an included script. Re-running the 7B dimension-restriction experiment from the repository alone reproduced the reported $k = 16$ native NLL change, KL, annotation AUROC and all retention decisions. Appendix A.1 maps each main-text claim to its supporting appendix material, and Appendix K lists the repository contents and requirements.

REFERENCES

Arc Institute. Evo 2 released model configurations. https://github.com/ArcInstitute/evo2/tree/v0.5.0/evo2/configs, 2026a. Zero-indexed attention schedules in evo2-7b-1m.yml and evo2-40b-1m.yml; accessed 26 September 2026.

Arc Institute. Evo 2 v0.5.0 release notes: Evo 2 20B via model surgery of Evo 2 40B. https://github.com/ArcInstitute/evo2/releases/tag/v0.5.0, February 2026b. Logit-lens analysis; later layers show abnormally high activation norms while contributing minimally to predictive cross-entropy.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, et al. Advancing regulatory variant effect prediction with alphagenome. *Nature*, 649(8099):1206–1218, 2026.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Gonzalo Benegas, Carlos Albors, Alan J Aw, Chengzhong Ye, and Yun S Song. A dna language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, 43(12):1960–1965, 2025.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A Gonzalez, et al. Genome modelling and design across all domains of life with evo 2. *Nature*, 652(8112):1349–1361, 2026. doi: 10.1038/s41586-026-10176-5.

CladeTeam. CENO-1B-base model card. https://huggingface.co/CladeTeam/CENO-1B-base, 2026. Base pretraining checkpoint; accessed 25 September 2026.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P De Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature methods*, 22(2):287–297, 2025.

Gregory M Findlay, Riza M Daza, Beth Martin, Melissa D Zhang, Anh P Leith, Molly Gasperini, Joseph D Janizek, Xingfan Huang, Lea M Starita, and Jay Shendure. Accurate classification of brca1 variants with saturation genome editing. *Nature*, 562(7726):217–222, 2018.

Adam Frankish, Sílvia Carbonell-Sala, Mark Diekhans, Irwin Jungreis, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Carme Arnan, If Barnes, et al. Gencode: reference annotation for the human and mouse genomes in 2023. *Nucleic acids research*, 51(D1):D942–D949, 2023.

Keisuke Hirota, Koichi Higashi, Ken Kurokawa, and Takuji Yamada. Benchmarking long-context genome language models on biosynthetic gene clusters. *bioRxiv*, pp. 2026–05, 2026.

Jerome Ku, Eric Nguyen, David W Romero, Garyk Brixi, Brandon Yang, Anton Vorontsov, Ali Taghibakhshi, Amy X Lu, Dave P Burke, Greg Brockman, et al. Systems and algorithms for convolutional multi-hybrid language models at scale. *arXiv preprint arXiv:2503.01868*, 2025.

Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, et al. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1):D1062–D1067, 2018.

Mingqian Ma, Yucheng Wu, Xin Chen, Feifei Jiang, Peijun Lin, Dongxin Ye, Yidi Sun, Yijing Zhang, Tianqiong Shi, Yu Zhao, et al. Ceno: A genome-scale world model for evolutionary sequence interpretation and programmable regulatory design. *bioRxiv*, pp. 2026–07, 2026.

Frederikke Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. Bend: Benchmarking dna language models on biologically meaningful tasks. In *International Conference on Learning Representations*, volume 2024, pp. 15246–15281, 2024.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201, 2023.

Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, 2010.

Enrique Queipo-de Llano, Alvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin. In *International Conference on Learning Representations*, volume 2026, pp. 13911–13937, 2026.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *Proceedings of machine learning research*, 235:43632, 2024.

Valerie A Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A Kitts, Terence D Murphy, Kim D Pruitt, Françoise Thibaud-Nissen, Derek Albracht, et al. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research*, 27(5):849–864, 2017.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Ziqi Tang, Nirali Somia, Yiyang Yu, and Peter K Koo. Evaluating the representational power of pre-trained dna language models for regulatory genomics. *Genome Biology*, 26(1):203, 2025.

Haoqi Wang, Tong Zhang, and Mathieu Salzmann. Demystifying singular defects in large language models. *arXiv preprint arXiv:2502.07004*, 2025.

APPENDIX CONTENTS

The appendices follow the order of the main-text claims. Panel definitions and shared measurement conventions come first; reproduction details come last.

Methods: shared definitions and panels

A Setup, conventions and the claim–evidence map 14
A.1 Claim–evidence map 14
A.2 Models, stored states and index conventions 14
A.3 Panels, labels and sampling 15
A.4 Probe protocol, estimands and uncertainty 16
A.5 Geometry and alignment definitions 17

Section 3.1: readout changes and controls

B Readouts at the transition 18
B.1 Full-depth trajectories 18
B.2 Local steepening and later declines 21
B.3 Probe-protocol sensitivity 21
B.4 Validation-stopped nonlinear probes 22
B.5 Identical positions 24
B.6 Context length, 6 kb versus 24 kb 25

C Annotation within similar predictive contexts 26

Sections 3.2–3.5: dimension, geometry, reconstruction and intervention

D Dimension restriction and native function 29

E Staged geometry 34
E.1 Canonical geometry, late inert blocks and rounding 34
E.2 Input composition 35
E.3 Arithmetic precision controls 35

F CENO recovery and model scope 37
F.1 CENO recovery and final-stage decomposition 37
F.2 HyenaDNA 39
F.3 Bidirectional screens: Caduceus-PS and NT-v2 39
F.4 Runtime validation and comparison limits 40

G Earlier-state reconstruction 40

H Native use of the transition 42
H.1 Natural-update attenuation and annotation recovery 42
H.2 Branch ablation 44
H.3 Branch decomposition and persistent low final rank 45
H.4 Dominant and residual update components 46

Section 4: biological readout

I Extraction depth for biological readout 46
I.1 Paired annotation readout and the chr18 replication 47
I.2 BEND gene-finding subset 48
I.3 Conservation 49
I.4 Extraction-state choice and nearby-layer sensitivity 49

J Variant-difference features 50
J.1 Panels, features and grouping 50
J.2 Comparison with the published GPN-MSA predictor 51

Reproduction

K Reproducibility, seeds and compute 52

A SETUP, CONVENTIONS AND THE CLAIM–EVIDENCE MAP

This appendix collects the conventions used throughout. Appendix A.1 maps each main-text claim to its supporting material; the remaining subsections define the stored states, the panels and labels, the probe protocol with its uncertainty scope, and the geometric statistics.

A.1 CLAIM–EVIDENCE MAP

Table A1 lists the claims of the main text in order, the appendix material that supports each one, and the panel on which that evidence was measured. Appendices B to H support Section 3; Appendices I and J support Section 4. Reproduction details and random seeds are collected in Appendix K.

Table A1: Supporting evidence in main-text order. D and C2 are disjoint 32-window chr22 panels; analyses on the same panel are not independent replications.

Main-text claim	Supporting appendix	Panel
Opposing annotation and next-base changes, with local steepening of annotation loss (§3.1)	B.1–B.2; probe, shared-position and context controls in B.3–B.6	Original chr22; D
Annotation decline persists within similar predictive contexts and on new inputs (§3.1)	C	D and C2
Sixteen leading transition-state directions retain native prediction while annotation falls (§3.2)	D	D; separate basis windows
Expansion, minimum rank and strongest alignment occur at different stages (§3.3)	E.1; arithmetic controls in E.3	Canonical chr22; precision panels
CENO recovers annotation while rank remains low; model profiles differ (§3.3, §5)	F.1; model-scope comparisons in F.2–F.4	Model-specific chr22 rows
Earlier-state structure remains partly recoverable after transition (§3.4)	G	chr22 and chr17
Tested attenuations do not jointly retain native prediction and recover final annotation (§3.5)	H.1; branch and component removal in H.2–H.4	D selects doses; C2 confirms; separate ablation panel
Earlier states improve annotation, gene finding and conservation; nearby depths are competitive (§4.1)	I	Development; retrospective chr19; new chr18; BEND subset
Variant features prefer different states; the published GPN-MSA predictor scores higher (§4.2)	J	chr17 panels; BRCA1

A.2 MODELS, STORED STATES AND INDEX CONVENTIONS

Architecture details. Evo 2 7B/40B have 32/50 StripedHyena 2 blocks of width 4096/8192. Each block's sequence mixer is attention or an input-dependent Hyena convolution with short, medium or long range; these types define the operator classes used in the branch controls. The released configurations place attention at zero-indexed blocks $\{3, 10, 17, 24, 31\}$ in 7B and $\{3, 10, 17, 24, 31, 35, 42, 49\}$ in 40B (Arc Institute, 2026a; Brixi et al., 2026; Ku et al., 2025). Only the 7B schedule has a uniform seven-block spacing. Architecture and training vary jointly across the retained models (Table A28).

Stored states and physical blocks. We compare computational stages: a block output, the input and output of the final block, and the final-normalised embedding. Stored b-indices are cache identifiers rather than a shared physical depth across models. Table A2 makes their correspondence explicit; cross-model depth plots use the fraction of physical blocks completed and display every final-normalised endpoint as `norm`.

The fixed Evo 2 transition states are b28/b21, with immediate pre-transition states b27/b20 (7B/40B); selected extraction states b24/b17 are earlier. CENO's comparison transition at cached b36 is produced by physical block 35. In Figure A5, CENO's pre-expansion, expansion and final-block

Table A2: Stored indices and physical computation. Physical block indices start at zero. The CENO replay records its final-block output separately from cached b38.

Model	Unnormalised states	Final endpoint
Evo 2	bi: physical block i output	Separate `norm` after the last block
CENO	b0: embedding; bi: physical block $i-1$ output for $1 \le i < L$	bL: last block followed by final norm ($L = 38$)
HyenaDNA	b0: embedding; b1–b8: physical blocks 0–7	b9: separate final LayerNorm
CENO final stages	b37: final-block input; X: physical block 37 output	b38: final RMSNorm output

input are cached b35, b36 and b37; the final-block output is the separately measured X, and final norm is cached b38. Appendix F.1 separates the effects of RMS scaling and learned coordinate gains. Where depth is the fraction of physical blocks completed, Evo 2 b28/b21 correspond to $29/32$ and $22/50$; an unnormalised HF block output bi corresponds to i/L. Final normalisation is reported separately from residual-block outputs.

Checkpoints for cross-model comparisons. We compare Evo 2 with two other genomic causal language models using the following pretrained checkpoints (Appendix F). CENO 1B uses the stage-2 base checkpoint `CladeTeam/CENO-1B-base` (Ma et al., 2026; CladeTeam, 2026); HyenaDNA uses `LongSafari/hyenadna-medium-160k-seqlen-hf` (Nguyen et al., 2023).

A.3 Panels, labels and sampling

Table A3: Panels and their roles. Extraction-state choices transfer to evaluation panels; probes are fitted separately within each. Analyses sharing a panel are not independent replications; the two 32-window chr22 panels use disjoint inputs.

Panel	Main role	Evaluation status
Original chr22/chr17	Depth profiles, reconstruction and extraction-state/offset selection	Development
12 chr22 basis windows	Fit state PCA and raw-update SVD bases	Separate from confirmation inputs
32-window chr22 (D)	Dimension/component interventions, probe/precision controls and natural-update dose selection	Disjoint from basis inputs and original chr22 windows; reused across analyses
Second 32-window chr22 (C2)	Intact and prediction-conditioned readout replication; natural-update confirmation	Inputs and size fixed before forward evaluation; disjoint from original, basis, pilot and D inputs
chr19	Compare fixed extraction states	Retrospective evaluation
New 7B chr18	Repeat fixed-state readout comparison	Panel fixed before outcomes; chromosome previously used

Matching is specified per assay and model: within-assay state comparisons keep sampled positions and folds fixed, whereas shared-position comparisons additionally match the named readouts. Stored row identities retain the input window and within-window position, so the same genomic coordinate in different contexts remains distinguishable.

The common comparison uses four binary tasks: coding exon, splice donor, splice acceptor and intergenic, each against intron. Additional 7B extraction analyses include $5'$UTR and $3'$UTR, giving six tasks; the 40B extraction analysis uses the common four. CENO's classwise appendix also reports UTR tasks, while its cross-model mean uses the common four. Thus the ten-cell extraction mean includes six 7B and four 40B results; the transition comparison uses the same four targets at both scales.

Window selection precedes position sampling. The original chr22/chr17 development panels select windows by annotations, not uniformly along the chromosome. Candidate windows are 6000

bases long on a 3000-base grid, with at most 1% ambiguous N bases. Using annotation counts in each window's central 3000 bases, the selector takes the top 14 windows for each of six categories: coding exon, intron, 5′UTR, 3′UTR, splice and intergenic. It unions these selections and fills to 100 with windows ranked by the number of categories present; ties retain the input window order. For both chromosomes, the six lists yield 84 distinct windows and the diversity step adds 16; the seeded random-fill fallback is unused. Chr19 uses a different rule: 6000-base windows on a nonoverlapping 6000-base grid must contain at least one splice-labelled position and at most 1% N bases; 100 eligible windows are sampled without replacement . The newly specified chr18 panel uses disjoint windows anchored on annotation categories (Appendix I.1). Thus none of these panels estimates readout or geometry under uniform whole-genome sampling. All focal inputs use the reference sequence in increasing coordinate order; transcript strand changes annotation labels, not input orientation.

Labels and position sampling. The source genome is GRCh38 (hg38) (Schneider et al., 2017) and region labels are derived from GENCODE v44 (Frankish et al., 2023). Different transcript isoforms can assign different regions to the same genomic position, and splice-boundary labels overlap intronic regions. A priority rule converts these overlapping annotations to one label per position. The main chr22 panel contains 100 windows of 6000 bases. Within each window, extraction retains all splice-labelled positions and samples 10% of coding-exon positions and 3.5% each of intron and intergenic positions; UTR-inclusive extractions additionally sample 25% of 5′UTR and 10% of 3′UTR positions. Each nonempty label receives at least one position. The probe-specific row caps below are applied after this extraction sampling. Overlap assignment uses, in increasing priority, intergenic, intron, CDS, 5′UTR, 3′UTR, splice donor and splice acceptor. Splice classes use the two intronic bases adjacent to exon boundaries with strand-aware coordinates. The stored "intergenic" code is the default for bases not assigned any of the retained labels, rather than a complement of annotated gene intervals. For example, noncoding exonic bases without a retained label can remain in this background. The intergenic task therefore contrasts this operational background, including some genic positions, with intron-labelled bases. Stored labels, windows and extraction row order are retained. In each binary task, the named class is positive and intron is negative. Each class is capped at 4000 randomly selected rows; all layers receive exactly the same selected rows.

Coordinate mapping and sampling. The reconstructed coordinates reproduce stored labels and window order exactly on all eight core cache panels. Coordinates are reconstructed from the extraction recipes because the original caches lack separately saved absolute positions; this check establishes internal consistency. The replay uses the frozen random seeds, label-selection order and input intervals. Canonical geometry uses one fixed random subset of 2048 rows without replacement, shared across states within each extraction. These rows inherit annotation-enriched sampling, with equal weight per selected position rather than per window.

Token and coordinate mapping. Input intervals are half-open throughout. A state is associated with the genomic position of the token that produced it, and the label of that position is taken from the annotation described in Appendix A.3. The focused models use single-base tokenisation. CENO's tokeniser was replayed to confirm one-base spans before its next-base targets were built (Appendix B.1). Variant features are indexed by the variant position in the reference and alternate inputs rather than by a single state index (Appendix J.1), and BEND's splice classes each carry their own offset and strand convention (Appendix I.2).

Sampling and precision. Early-layer caches use FP16 storage; late overflowing states were re-extracted and stored in FP32 with row identity checked. Storage precision is distinct from model arithmetic. Evo 2 7B uses native bf16 branch computation; 40B uses its Transformer Engine FP8 projection path. The 40B precision controls retain native branch-computation dtypes.

A.4 Probe protocol, estimands and uncertainty

Grouped folds. Input intervals are half-open. Two windows are connected when their intervals overlap, and transitive overlap defines a group. The chr22 100-window panel has 66 connected groups; chr17 merges windows as well, whereas the chr19 windows do not overlap. Grouping by window ID alone does not guarantee disjoint input sequence. Five-fold GroupKFold is used unless a training or evaluation fold has only one binary class. That occurs for chr22 intergenic; a fixed

StratifiedGroupKFold with fixed shuffling supplies valid folds. The fallback depends only on label support and uses the same folds at every layer, independently of model scores.

Fitting and scoring. Coordinate standardisation, PCA256 and logistic regression are fitted within each training fold. The score is the unweighted mean of the five held-out AUROCs, not the AUROC of pooled predictions; different folds can have different calibration. Predictions, row IDs, positive labels, coordinates, groups and fold IDs are stored. All development block outputs were refitted under this protocol; chr19 uses the fixed extraction states and all selected comparison states.

Probe orientation is fixed by training labels; held-out AUROCs below 0.5 are retained.

Estimand and conditional uncertainty. For a chromosome with C model–task cells, the extraction contrast is

$$\widehat{\delta} = \frac{1}{C}\sum_{c=1}^{C}\left[\frac{1}{5}\sum_{f=1}^{5}\left(\mathrm{AUROC}_{c,f}^{\mathrm{selected}} - \mathrm{AUROC}_{c,f}^{\mathrm{final}}\right)\right].$$

Each cell and fold has equal weight; $C = 10$ for the chr19 extraction mean. The interval describes uncertainty under reweighting genomic groups, conditional on the fitted predictions and selected layers. It does not include probe-refitting or selector uncertainty. The bootstrap has three steps:

1. In each of 2000 draws, assign independent Exponential(1) weights to the union of genomic groups across scales. Every row inherits its group's weight, shared across tasks, states and scales.
2. Recompute weighted AUROC within each fixed fold, then the fold and cell averages above.
3. Retain the paired difference; the 2.5th and 97.5th percentiles give the interval.

Chromosomes use independent draws. The focal four-task transition analysis uses the same aggregation logic. Random seeds for all reported assays are listed in Table A48. Overlapping intervals are not a test of equivalence.

Direct intermediate-state decoding. Logit-lens analysis of Evo 2 (Arc Institute, 2026b) applies the output head to intermediate states; the tuned lens (Belrose et al., 2023) learns a decoder for each state. These differ from our native prediction measurements, which rerun the remaining pretrained computation. They provide methodological context here, not an additional lens experiment.

Probe capacity. The primary protocol fixes preprocessing, projection dimension and regularisation across states, with parameters fitted independently within each training fold. Appendix B.3 varies PCA dimension, removes projection, selects regularisation within training folds and evaluates positionwise RMS scaling and a fixed-budget nonlinear probe. These controls test robustness of the transition/pre-transition and final/pre-transition contrasts. The separate adjacent-change analysis tests local steepening (Appendix B.2).

A.5 Geometry and alignment definitions

State and geometric notation. At block index ℓ and input position t, the causal state $h_{\ell,t}$ represents input bases through x_t; annotation concerns t and the next-base target is x_{t+1}. On the same position sample at each state, define mean state norm, adjacent norm ratio and standardised effective rank as

$$n_\ell = \mathbb{E}_t\|h_{\ell,t}\|_2, \qquad \rho_\ell = n_\ell/n_{\ell-1}, \qquad R_\ell = \exp\left(-\sum_i p_{\ell,i}\log p_{\ell,i}\right). \tag{1}$$

Here $p_{\ell,i}$ are normalised covariance eigenvalues after centring and coordinate-wise standardisation across positions. If k directions share variance equally, $R_\ell = k$; if one dominates, effective rank approaches one. The main text names these measurements in words; the formal symbols are used only in the appendix. The primary rank applies the entropy effective rank of Roy & Vetterli (2007) to the coordinate-standardised covariance spectrum. It describes variation across genomic positions; compression-valley analyses instead emphasise token-matrix spectra dominated by the beginning-of-sequence token (Queipo-de Llano et al., 2026). Neither statistic counts biological features.

The realised update and its scale relative to the incoming state are

$$\Delta_{\ell,t} = h_{\ell,t} - h_{\ell-1,t}, \qquad \omega_\ell = \mathbb{E}_t \|\Delta_{\ell,t}\|_2 / n_{\ell-1}. \tag{2}$$

Scale ratios and the realised update. For ρ_ℓ as defined in Equation (1) and ω_ℓ as defined in Equation (2), both ratios of means over the same rows, the triangle inequality applied inside the expectation gives

$$|\rho_\ell - 1| \;\le\; \omega_\ell \;\le\; \rho_\ell + 1. \tag{3}$$

Both bounds are checked numerically at every tabulated state. Because ρ_ℓ and ω_ℓ are ratios of means rather than means of ratios, neither equals the corresponding per-position average. A realised Δ_ℓ can also differ from the branch tensor computed before addition, because the addition is performed in the runtime dtype and part of a small write can be lost to rounding. Branch writes measured before addition, as in Table A37, are therefore not interchangeable with realised state differences, and Appendix E.3 quantifies the discrepancy at the blocks where it matters.

Spectral definitions and their edge cases. For a matrix X, the raw and centred effective ranks use eigenvalues of $X^\top X$ and $X_c^\top X_c$, respectively. Standardised state rank replaces each nonconstant column of X_c by its population-standard-deviation normalisation. With 2048 centred rows and hidden width d, effective rank is bounded by $\min(2047, d)$. An exactly constant column is zeroed; there is no additive variance floor, so a coordinate that never varies across positions cannot contribute variance through a floating-point mean residual. The entropy uses normalised nonnegative eigenvalues, with zero eigenvalues omitted. A zero matrix is assigned rank zero. Computation uses the smaller Gram matrix when appropriate, global rescaling before decomposition and FP64 analysis. Nonfinite inputs raise an error instead of being replaced.

Output basis. Let Q be an orthonormal basis for the span of the four nucleotide head rows after subtracting their mean. For the centred state matrix $H_{\ell,c}$, output-contrast alignment is $A_\ell = \|H_{\ell,c}Q\|_F^2 / \|H_{\ell,c}\|_F^2$. The measured contrast subspace has rank 3, with random-subspace expectation $3/d$ for hidden width d: 0.073% in Evo 2 7B and 0.037% in 40B. Alignment retains the original feature scales, whereas standardised rank equalises them. The output basis uses the runtime head weights: `embedding_layer.weight` (tied) in Evo 2 and `lm_head.weight` in CENO. A relative singular-value threshold 10^{-6} gives rank three after subtracting the mean ACGT row. This basis does not represent the entire softmax or constrain the residual to rank three, and is not fitted to maximise alignment.

Before final normalisation, alignment measures the fraction of positional variance already in the head's contrast subspace. Subsequent RMSNorm applies positionwise rescaling and learned coordinate gains, so this geometric fraction differs from a direct contribution to the logits.

B Readouts at the transition

This appendix supports Section 3.1. It gives the full-depth readout trajectories, the local-steepening contrast, sensitivity to the probe protocol and to a validation-stopped nonlinear decoder, and the shared-position and context-length replications.

B.1 Full-depth trajectories under connected-window refitting

Both Evo 2 transitions improve fitted next-base accuracy and NLL; CENO's transition degrades both. The following assay tests access to next-base targets from each state, separately from the pretrained output head. It uses 8000 fixed rows per model sampled without replacement from the eligible chr22 cache. Targets are the hg38 base at position $p + 1$ for a state at p; both bases must be ACGT and the target must remain inside the original input window. The annotation-enriched cache population is inherited. Coordinates use the verified recipe replay of Appendix A.3. CENO's tokeniser additionally verifies one-base spans on all $600{,}000$ source bases. All three samples cover 100 windows in 66 overlap-connected groups.

Five GroupKFold folds are fixed across states within each model. Training-fold StandardScaler, PCA256 and multinomial logistic regression ($C = 1$) predict A/C/G/T. Frequency baselines use

training-fold base frequencies and the corresponding majority base. Accuracy and NLL are unweighted means of held-out fold scores. Their 2000 paired exponential-weight cluster draws condition on the fitted predictions. The intervals exclude sampling, fitting and model-training uncertainty. Fitted-probe scores measure accessibility under the learned decoder. Next-base and binary annotation assays use different samples and can use different folds, supporting comparisons across depth at the assay level.

All next-base fits converged.

The full-depth annotation curves in Figure A1 place the focal contrasts within each task's trajectory and also supply the extraction-calibration scores.

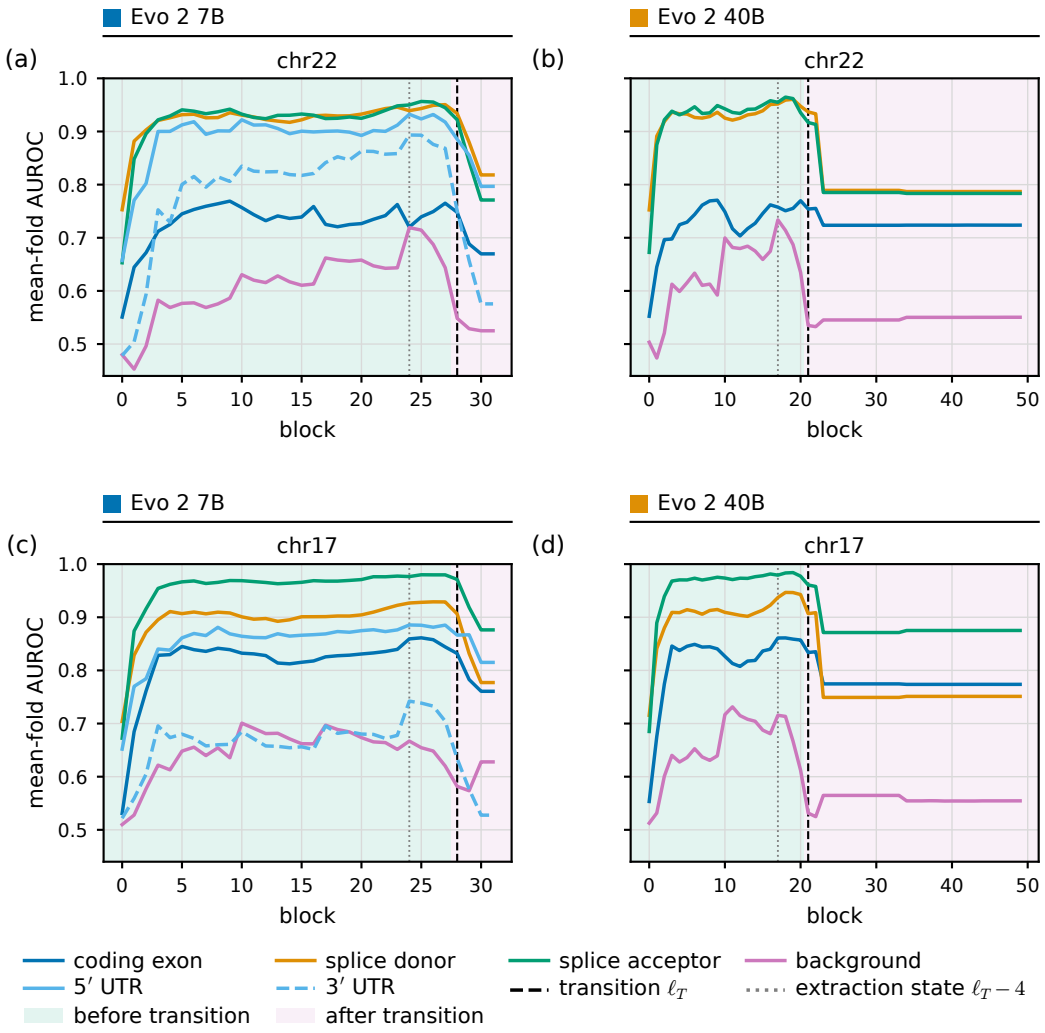

Figure A1: Development-block annotation AUROCs, averaged over five connected-window folds. Each curve uses the same rows and folds across depth within its task; 7B has six tasks and 40B four. Background is the stored `intergenic` label. The black dashed line marks the fixed transition and the grey dotted line the extraction state four blocks earlier. These curves define the calibration objective; chr19 is absent.

The final states improve next-base decodability over immediate pre-transition states in all three models, although the intervening trajectories need not be monotonic: 40B b23 has higher NLL than b21 despite similar accuracy. The local-depth extension retains these focal predictions.

Table A4: Fitted next-base decodability on chr22. All focal states within a model share 8,000 sampled positions, five overlap-connected folds, and the train-fold scaler–PCA256–multinomial logistic protocol. Accuracy and four-class NLL (nats) are means across held-out folds; brackets show conditional paired-cluster intervals. The frequency baseline is fitted on each training fold. These are probe metrics, not native model-head metrics.

Model	State	Accuracy [95% interval]	NLL [95% interval]
Evo 2 7B	train-frequency	0.273 [0.259, 0.290]	1.383 [1.377, 1.389]
Evo 2 7B	**b24**	0.485 [0.460, 0.504]	1.186 [1.156, 1.219]
Evo 2 7B	b27	0.553 [0.521, 0.579]	1.050 [1.002, 1.104]
Evo 2 7B	b28	0.572 [0.539, 0.598]	0.975 [0.924, 1.035]
Evo 2 7B	b30	0.590 [0.559, 0.617]	0.947 [0.901, 1.002]
Evo 2 7B	final norm	0.592 [0.561, 0.619]	0.920 [0.870, 0.976]
Evo 2 40B	train-frequency	0.266 [0.252, 0.287]	1.385 [1.377, 1.391]
Evo 2 40B	**b17**	0.560 [0.538, 0.581]	1.088 [1.048, 1.130]
Evo 2 40B	b20	0.672 [0.637, 0.711]	0.820 [0.738, 0.892]
Evo 2 40B	b21	0.700 [0.666, 0.737]	0.739 [0.655, 0.812]
Evo 2 40B	b23	0.702 [0.667, 0.739]	0.803 [0.721, 0.874]
Evo 2 40B	final norm	0.702 [0.669, 0.739]	0.715 [0.634, 0.786]
CENO	train-frequency	0.276 [0.260, 0.295]	1.384 [1.379, 1.389]
CENO	b35	0.481 [0.461, 0.501]	1.166 [1.131, 1.198]
CENO	b36	0.308 [0.295, 0.323]	1.365 [1.357, 1.372]
CENO	b37	0.305 [0.293, 0.319]	1.363 [1.355, 1.370]
CENO	b38	0.572 [0.546, 0.600]	0.999 [0.951, 1.048]

Table A5: Final-state next-base contrasts in Evo 2, and transition and recovery contrasts in CENO. Adjacent Evo 2 transition changes are reported in Table A8. Positive accuracy and negative NLL indicate improvement. Intervals use 2,000 common Exponential(1) weights over 66 connected input-context groups and condition on fitted out-of-fold probes.

Model	Contrast	Δ accuracy [95% interval]	Δ NLL [95% interval]
7B	norm − b27	+0.039 [+0.030, +0.048]	-0.131 [-0.145, -0.115]
7B	norm − b24	+0.107 [+0.091, +0.126]	-0.267 [-0.304, -0.229]
40B	norm − b20	+0.030 [+0.020, +0.039]	-0.105 [-0.117, -0.091]
40B	norm − b17	+0.142 [+0.119, +0.171]	-0.374 [-0.431, -0.328]
CENO	b36 − b35	-0.174 [-0.200, -0.147]	+0.199 [+0.164, +0.236]
CENO	b38 − b35	+0.090 [+0.071, +0.111]	-0.167 [-0.205, -0.129]
CENO	b38 − b36	+0.264 [+0.234, +0.292]	-0.366 [-0.415, -0.316]

For the annotation contrasts in Table A7, we reanalyse the existing OOF predictions for coding exon, splice donor, splice acceptor and intergenic versus intron. Each model's fixed states share rows, labels, groups and folds within each task. The mean of four task-specific mean-fold AUROCs uses 2000 cluster draws, with weights shared across tasks and states within the model. CENO's final shared-class annotation readout recovers relative to transition but remains below its immediate pre-transition value.

Table A6: Paired changes in annotation AUROC on chr22, averaged over four binary annotation tasks. Pre denotes the immediate pre-transition state; intervals are conditional on the fitted OOF predictions.

Model	Transition − pre	Final − pre	Final − transition
Evo 2 7B	-0.038 $[-0.048, -0.025]$	-0.117 $[-0.139, -0.095]$	-0.079 $[-0.098, -0.062]$
Evo 2 40B	-0.036 $[-0.047, -0.027]$	-0.108 $[-0.132, -0.092]$	-0.072 $[-0.091, -0.058]$
CENO	-0.149 $[-0.164, -0.133]$	-0.035 $[-0.046, -0.025]$	$+0.114$ $[+0.098, +0.129]$

The UTR-inclusive 7B panel. The 7B cache carries both UTR classes, so the same transition can also be averaged over six binary annotation tasks on that panel. The six-task chr22 means are 0.859 at b24, 0.848 at b27, 0.798 at b28 and 0.707 at final normalisation. Cross-scale comparisons use the four binary annotation tasks in Table A6.

B.2 Local steepening and later declines

For score s_ℓ, let $\delta_\ell = s_\ell - s_{\ell-1}$. At the fixed transition ℓ_T,

$$\delta_{\text{excess}} = \delta_{\ell_T} - \frac{1}{3}\sum_{j=1}^{3} \delta_{\ell_T - j}. \tag{4}$$

Table A7: Local readout changes at the fixed b28/b21 expansions. Relative to the preceding three transitions, annotation decline steepens; next-base accuracy improves without detectable acceleration. Brackets give 95% paired intervals.

Readout	Before → after	Change	Excess change
■ *Evo 2 7B (b27→b28)*			
Annotation AUROC	0.826 → 0.788	−0.038 [−0.048, −0.025]	−0.036 [−0.047, −0.022]
Next-base accuracy	0.553 → 0.572	+0.019 [+0.009, +0.028]	−0.004 [−0.016, +0.008]
■ *Evo 2 40B (b20→b21)*			
Annotation AUROC	0.822 → 0.786	−0.036 [−0.047, −0.027]	−0.027 [−0.038, −0.017]
Next-base accuracy	0.672 → 0.700	+0.027 [+0.019, +0.035]	−0.010 [−0.024, +0.002]

The transition annotation decline is steeper than the preceding mean, while next-base accuracy gains are not detectably larger. The local next-base analysis uses the same 8000 positions, targets, overlap groups and preprocessing at all states; logistic fits use $C = 1$ and a 20,000-iteration cap and all converge. The comparison window and excess-change contrast were fixed before these fits, with focal results already known. The preceding states b24–b27 (7B) and b17–b20 (40B) define three adjacent increments. Annotation contrasts are retrospective reanalyses of these same ranges; robustness to alternative window widths has not been evaluated. Shared genomic-group weights pair all state contrasts, conditional on fitted predictions.

Accuracy improves at the transition in both models, but the estimated gain is smaller than the mean gain across the three preceding transitions. The differences are -0.004 $[-0.016, +0.008]$ and -0.010 $[-0.024, +0.002]$ for 7B and 40B. NLL gives a scale-dependent result: the transition reduction exceeds the preceding average reduction in 7B by 0.030 nats, whereas the 40B interval includes zero. Excess predictive gains therefore vary by metric and scale. Reanalysis of the stored annotation predictions instead gives transition-minus-prior-average changes of -0.036 $[-0.047, -0.022]$ in 7B and -0.027 $[-0.038, -0.017]$ in 40B. The local contrast therefore combines a steeper annotation decline with an accuracy improvement already present before the boundary. Later annotation decreases are also substantial: b28→b29 at 7B and b22→b23 at 40B exceed the respective transition drops by point estimate.

Table A7 reports the paired transition and excess contrasts.

B.3 Probe-protocol sensitivity

The annotation decline persists across the ten original protocols at both scales. This sensitivity analysis reuses the original chr22 rows and overlap-connected outer folds at the immediate pre-transition, transition and final states. Fixed-C linear probes compare 64, 128, 256 and 512 principal components with unprojected features. Each training subset supplies one full-SVD factorisation, whose leading components define the lower-dimensional arms. A separate analysis selects $C \in \{0.01, 0.1, 1, 10\}$ using three grouped folds within each outer training fold, with standardisation and PCA refitted inside every inner fold. Additional fixed arms use positionwise RMS normalisation before PCA256 and a 64-unit ReLU MLP after PCA256. The MLP uses a fixed training budget and no validation-based early stopping.

Table A8: Adjacent next-base changes on the original chr22 rows and genomic folds. Accuracy and NLL changes use the later state minus the earlier state. The last row for each model subtracts the mean change over the three immediately preceding transitions from the transition change. Brackets give conditional paired 95% genomic-cluster intervals.

Model	Transition	Accuracy change	NLL change (nats)
Evo 2 7B	b24→b25	+0.029 [+0.018, +0.040]	−0.051 [−0.070, −0.033]
	b25→b26	+0.022 [+0.014, +0.028]	−0.042 [−0.055, −0.028]
	b26→b27	+0.018 [+0.010, +0.027]	−0.044 [−0.055, −0.032]
	b27→b28	+0.019 [+0.009, +0.028]	−0.075 [−0.089, −0.059]
	b28→b29	+0.015 [+0.008, +0.024]	−0.045 [−0.054, −0.037]
	b29→b30	+0.003 [−0.003, +0.010]	+0.016 [+0.010, +0.022]
	b30→b31	+0.000 [+0.000, +0.000]	+0.000 [+0.000, +0.000]
	Excess transition change	−0.004 [−0.016, +0.008]	−0.030 [−0.046, −0.011]
Evo 2 40B	b17→b18	+0.042 [+0.030, +0.056]	−0.100 [−0.129, −0.076]
	b18→b19	+0.045 [+0.032, +0.059]	−0.102 [−0.130, −0.079]
	b19→b20	+0.025 [+0.017, +0.036]	−0.066 [−0.083, −0.052]
	b20→b21	+0.027 [+0.019, +0.035]	−0.081 [−0.094, −0.067]
	b21→b22	−0.006 [−0.011, −0.001]	+0.010 [+0.005, +0.014]
	b22→b23	+0.008 [+0.001, +0.014]	+0.053 [+0.042, +0.064]
	b23→b24	+0.000 [+0.000, +0.000]	+0.000 [+0.000, +0.000]
	Excess transition change	−0.010 [−0.024, +0.002]	+0.008 [−0.014, +0.033]

The transition-minus-immediate-pre-transition and final-minus-immediate-pre-transition intervals are below zero in every tested setting at both scales (Table A9). This includes unprojected features, training-selected regularisation and positionwise RMS scaling. These controls evaluate transition and final-state contrasts; Appendix B.2 evaluates local steepening under the primary protocol. The results characterise the tested decoder families and budgets. In particular, 65 of 120 MLP fits reach the 300-epoch cap. Appendix B.4 adds a separate validation-stopped analysis and reports both its primary protocol and a secondary scaling sensitivity.

Adopted fitting settings. Linear arms use $C = 1$ unless selected within training folds. Positionwise RMS scaling divides each row by its own root-mean-square coordinate value before the training-fold scaler; zero rows remain zero. New linear fits use Newton–CG for the same L_2-penalised logistic objective, stopping at a training-gradient infinity norm of 10^{-4} up to numerical rounding. The MLP has one 64-unit ReLU layer, Adam learning rate 10^{-3}, batch size 256, L_2 coefficient 10^{-3} and a 300-epoch cap. Its stopping rule uses training loss only. Outer-fold labels select neither the protocol nor the optimisation budget.

All linear probe-sensitivity fits meet the training-gradient tolerance.

B.4 Validation-stopped nonlinear annotation probes

We fit a nonlinear decoder on panel D to assess annotation accessibility with validation-based training duration. The earlier fixed-budget MLP (Appendix B.3) and this analysis use different panels and architectures. Panel D is the 32-window dimension-restriction panel, with the same four states, task rows and grouped folds as Appendix C. The reference is the no-PCA $C = 1$ logistic probe fitted on those rows, rather than the original PCA256 annotation assay.

Fitting and stopping. Each split fits coordinate standardisation and PCA1024 on its training rows. The decoder has 1024 inputs, 256 GELU hidden units and one logistic output. AdamW uses learning rate 10^{-3}, weight decay 10^{-4} and batch size 512. Validation uses the first of three GroupKFold splits within the outer training set; both standardisation and PCA are refitted inside this split. Training stops after 60 epochs without a validation-log-loss improvement greater than 10^{-5}, with a 2000-epoch cap. The selected best epoch count is then used to refit the decoder on the full outer training set. Outer held-out rows are used only for scoring. Four states, four tasks and five outer folds give 80 fits per model.

Table A9: Annotation-probe sensitivity on fixed chr22 rows and overlap-connected folds, averaged over four binary annotation tasks. Entries give AUROC changes and pointwise 95% paired genomic-cluster intervals conditional on fitted predictions and training-selected regularisation. New PCA arms use exact SVD; the original auto-PCA256 baseline is retained.

Evo 2 7B		
Probe	Transition − pre	Final − pre
Original PCA256, $C = 1$	-0.038 $[-0.048, -0.025]$	-0.117 $[-0.139, -0.095]$
PCA64, $C = 1$	-0.045 $[-0.061, -0.032]$	-0.099 $[-0.121, -0.076]$
PCA128, $C = 1$	-0.043 $[-0.055, -0.029]$	-0.112 $[-0.135, -0.090]$
PCA256, $C = 1$	-0.039 $[-0.048, -0.026]$	-0.118 $[-0.140, -0.095]$
PCA512, $C = 1$	-0.027 $[-0.037, -0.015]$	-0.116 $[-0.137, -0.095]$
No PCA, $C = 1$	-0.023 $[-0.033, -0.011]$	-0.116 $[-0.137, -0.095]$
PCA256, nested C	-0.041 $[-0.051, -0.028]$	-0.122 $[-0.144, -0.099]$
No PCA, nested C	-0.021 $[-0.028, -0.012]$	-0.131 $[-0.153, -0.107]$
Row RMS + PCA256, $C = 1$	-0.047 $[-0.056, -0.033]$	-0.119 $[-0.142, -0.096]$
PCA256 + MLP64	-0.037 $[-0.047, -0.028]$	-0.085 $[-0.103, -0.073]$

Evo 2 40B		
Probe	Transition − pre	Final − pre
Original PCA256, $C = 1$	-0.036 $[-0.047, -0.027]$	-0.108 $[-0.132, -0.092]$
PCA64, $C = 1$	-0.034 $[-0.047, -0.026]$	-0.079 $[-0.103, -0.065]$
PCA128, $C = 1$	-0.040 $[-0.051, -0.032]$	-0.100 $[-0.124, -0.085]$
PCA256, $C = 1$	-0.036 $[-0.047, -0.027]$	-0.108 $[-0.132, -0.092]$
PCA512, $C = 1$	-0.031 $[-0.042, -0.020]$	-0.111 $[-0.133, -0.094]$
No PCA, $C = 1$	-0.042 $[-0.056, -0.029]$	-0.130 $[-0.153, -0.112]$
PCA256, nested C	-0.036 $[-0.046, -0.027]$	-0.112 $[-0.136, -0.096]$
No PCA, nested C	-0.031 $[-0.041, -0.021]$	-0.138 $[-0.160, -0.120]$
Row RMS + PCA256, $C = 1$	-0.035 $[-0.048, -0.025]$	-0.106 $[-0.130, -0.091]$
PCA256 + MLP64	-0.034 $[-0.049, -0.022]$	-0.080 $[-0.100, -0.069]$

Every primary fit reaches the validation-stopping criterion before the cap (Table A13). The selected epoch medians are 6 and 7 in 7B/40B; 21 and 19 fits select epoch 1. A validation plateau establishes completion of this protocol, not optimality over nonlinear decoders. Scores and depth contrasts use the same 2000 paired genomic-group weights as the linear comparison, conditional on the fitted probes.

Table A10: Primary validation-stopped MLP: four-task AUROC and paired differences from the no-PCA linear reference on the same rows and folds.

State	Linear	Nonlinear	Nonlinear − linear
■ Evo 2 7B			
b27	0.7716 $[0.7382, 0.8153]$	0.7749 $[0.7430, 0.8198]$	$+0.0033$ $[-0.0052, +0.0133]$
b28	0.7444 $[0.7072, 0.7906]$	0.7351 $[0.7012, 0.7772]$	-0.0093 $[-0.0244, +0.0055]$
b30	0.6579 $[0.6137, 0.7031]$	0.6789 $[0.6364, 0.7196]$	$+0.0210$ $[+0.0054, +0.0332]$
Final	0.6813 $[0.6498, 0.7122]$	0.6928 $[0.6558, 0.7309]$	$+0.0115$ $[+0.0001, +0.0241]$
■ Evo 2 40B			
b20	0.7894 $[0.7539, 0.8367]$	0.7899 $[0.7583, 0.8318]$	$+0.0005$ $[-0.0114, +0.0116]$
b21	0.7761 $[0.7449, 0.8158]$	0.7477 $[0.7207, 0.7850]$	-0.0284 $[-0.0418, -0.0160]$
b23	0.6627 $[0.6302, 0.6990]$	0.6761 $[0.6489, 0.7132]$	$+0.0134$ $[+0.0001, +0.0336]$
Final	0.6759 $[0.6498, 0.7056]$	0.6933 $[0.6644, 0.7278]$	$+0.0174$ $[+0.0004, +0.0357]$

The primary MLP retains negative initial, subsequent and pre-to-final contrasts in both models. It also improves minimum-rank readout over the linear reference by $0.0210/0.0134$ and final readout by $0.0115/0.0174$ in 7B/40B. The final-minus-pre-transition gaps remain $-0.0822/-0.0965$. Thus decoder flexibility recovers some later accessibility without closing the depth gap under this

Table A11: Depth contrasts under both nonlinear protocols. Secondary scaling reverses the initial contrast while preserving the later and pre-to-final declines.

Contrast	*Evo 2 7B*	*Evo 2 40B*
Primary		
Pre → transition	-0.0398 $[-0.0585, -0.0267]$	-0.0422 $[-0.0661, -0.0225]$
Transition → minimum	-0.0562 $[-0.0821, -0.0422]$	-0.0716 $[-0.0880, -0.0573]$
Pre → final	-0.0822 $[-0.1172, -0.0615]$	-0.0965 $[-0.1268, -0.0762]$
Secondary: post-PCA scaling		
Pre → transition	$+0.0149$ $[+0.0004, +0.0269]$	$+0.0148$ $[+0.0007, +0.0298]$
Transition → minimum	-0.1196 $[-0.1489, -0.0946]$	-0.1264 $[-0.1586, -0.0981]$
Pre → final	-0.1146 $[-0.1469, -0.0930]$	-0.1390 $[-0.1812, -0.1047]$

protocol. Relative to the linear probe, the transition-to-minimum-rank decline is smaller by 0.0302 $[0.0077, 0.0507]$ in 7B and 0.0418 $[0.0263, 0.0631]$ in 40B. These are differences of depth contrasts, not the absolute gains at the final embedding.

Secondary coordinate-scaling sensitivity. An additional variant standardises the PCA coordinates on the training rows before the MLP. This variant was not part of the locked primary protocol and is reported separately. All 80 fits per model again meet validation stopping before the cap. Initial contrasts reverse to $+0.0149$ $[+0.0004, +0.0269]$ in 7B and $+0.0148$ $[+0.0007, +0.0298]$ in 40B. Subsequent and pre-to-final decreases persist. Its absolute macro AUROC is below the linear reference at every state (Table A12). The initial contrast is therefore sensitive to decoder preprocessing even though the later depth gap occurs under both variants.

Table A12: Secondary MLP with post-PCA coordinate standardisation: four-task AUROC and paired differences from the no-PCA linear reference on the same rows and folds.

State	Linear	Nonlinear	Nonlinear − linear
■ Evo 2 7B			
b27	0.7716 $[0.7382, 0.8153]$	0.7030 $[0.6723, 0.7417]$	-0.0686 $[-0.0836, -0.0557]$
b28	0.7444 $[0.7072, 0.7906]$	0.7180 $[0.6827, 0.7556]$	-0.0264 $[-0.0429, -0.0148]$
b30	0.6579 $[0.6137, 0.7031]$	0.5983 $[0.5749, 0.6216]$	-0.0596 $[-0.0853, -0.0353]$
Final	0.6813 $[0.6498, 0.7122]$	0.5885 $[0.5689, 0.6064]$	-0.0928 $[-0.1122, -0.0732]$
■ Evo 2 40B			
b20	0.7894 $[0.7539, 0.8367]$	0.7363 $[0.6987, 0.7803]$	-0.0531 $[-0.0673, -0.0427]$
b21	0.7761 $[0.7449, 0.8158]$	0.7511 $[0.7201, 0.7910]$	-0.0249 $[-0.0400, -0.0113]$
b23	0.6627 $[0.6302, 0.6990]$	0.6247 $[0.6004, 0.6550]$	-0.0380 $[-0.0527, -0.0226]$
Final	0.6759 $[0.6498, 0.7056]$	0.5973 $[0.5830, 0.6131]$	-0.0786 $[-0.0992, -0.0626]$

Table A13: Validation-stopping diagnostics. Selected epochs are reported as minimum/median/maximum; reaching a plateau is the specified stopping criterion.

Model, variant	Fits	Plateau	Cap	Epochs	Epoch 1
■ 7B, Primary	80	80	0	1/6/59	21
■ 7B, Post-PCA scaling	80	80	0	1/2.5/20	31
■ 40B, Primary	80	80	0	1/7/90	19
■ 40B, Post-PCA scaling	80	80	0	1/3/18	20

B.5 Annotation and prediction on identical positions

We repeat the local readout comparison with both targets defined on each sampled position. For each of the four annotation tasks, the binary annotation target and next-base A/C/G/T target use identical eligible rows, overlap-connected folds and training-fitted StandardScaler–PCA256 features.

Separate logistic classifiers use $C = 1$. Ineligible final positions and ambiguous bases are removed before fixing the panel. Scores average the five held-out folds and then the four tasks. The 7B 6/24 kb comparison uses the same rows and folds connected by overlapping 24 kb inputs; the 40B 6 kb comparison uses its original input groups. These annotation-enriched, task-specific next-base samples differ from the separate next-base panel in Table A4.

Table A14: Shared-position readout changes. Each entry gives a transition change and paired genomic-group 95% interval. Excess change subtracts the mean of the preceding three adjacent changes. Both classifiers share rows and training transforms within each task.

Model / context	Metric	Transition change	Excess change
■ 7B / 6 kb	Annotation AUROC	-0.030 $[-0.040, -0.020]$	-0.032 $[-0.042, -0.021]$
■ 7B / 6 kb	Next-base accuracy	$+0.011$ $[+0.004, +0.015]$	-0.009 $[-0.018, -0.003]$
■ 7B / 6 kb	Next-base NLL	-0.037 $[-0.044, -0.029]$	$+0.006$ $[-0.002, +0.015]$
■ 7B / 24 kb	Annotation AUROC	-0.044 $[-0.055, -0.031]$	-0.044 $[-0.055, -0.029]$
■ 7B / 24 kb	Next-base accuracy	$+0.015$ $[+0.008, +0.020]$	-0.004 $[-0.013, +0.002]$
■ 7B / 24 kb	Next-base NLL	-0.037 $[-0.044, -0.028]$	$+0.008$ $[-0.000, +0.018]$
■ 40B / 6 kb	Annotation AUROC	-0.037 $[-0.053, -0.025]$	-0.031 $[-0.047, -0.017]$
■ 40B / 6 kb	Next-base accuracy	$+0.021$ $[+0.016, +0.025]$	-0.011 $[-0.017, -0.005]$
■ 40B / 6 kb	Next-base NLL	-0.072 $[-0.078, -0.063]$	-0.006 $[-0.017, +0.006]$

Annotation AUROC decreases and next-base accuracy increases in all three shared-position comparisons. The excess annotation change is negative throughout. At 6 kb, the excess accuracy change is also negative at both scales: accuracy continues to improve across transition at a lower rate than over the preceding three transitions. At 24 kb, the excess accuracy interval spans zero. These comparisons reproduce the opposing readout changes without assigning the assays different sampled positions.

B.6 Context-length sensitivity at fixed genomic positions

The primary analysis uses 6 kb windows. To test sensitivity to additional left context, we repeat Evo 2 7B extraction on the same 100 chr22 windows after prepending 18 kb, giving 24 kb inputs with unchanged endpoints and readout positions. The checkpoint, provider configuration and tokenisation are identical at both lengths. Window coordinates, position samples, states, folds and analysis were fixed before these additional forward passes; the original short-context results were already known.

Overlapping 24 kb inputs form 60 connected groups. These groups define folds at both lengths, preventing added context from crossing training and evaluation folds. Each of the four binary annotation tasks uses the same 4000 positive and 4000 intron positions at both lengths; the separately sampled next-base assay uses 8000 positions. Folds are fixed separately per task, with the label-support fallback in Appendix A.4. Geometry uses 2048 fixed positions and the same norm-ratio and standardised-rank calculations as the primary analysis.

Probes are fitted independently at b24–b28 and the final embedding for each length. Training-fold StandardScaler and PCA(256) precede logistic regression with $C = 1$. The iteration caps are 3000 for binary annotation and 20000 for next-base prediction; binary fits may be completed to 20000 if a convergence warning occurs, without selecting settings from held-out scores. The local annotation contrast remains the b27→b28 change minus the mean of the three preceding adjacent changes. Shared group weights (2000 Exponential(1) draws) pair comparisons across states and lengths. Intervals condition on fitted probes. The altered overlap groups and next-base position sample make the 6 kb arm a fresh paired baseline, rather than a reuse of the primary trajectory's predictions.

All 300 fold fits completed without convergence warnings. The short-context extraction exactly reproduces the cached states at the checked positions. Both lengths retain the b28 expansion, followed by much lower rank at b30. At 24 kb the annotation transition change and its local excess both remain negative. Next-base probe NLL improves across the transition, whereas its accuracy interval includes zero (Table A15). Accuracy also improves less than over the preceding three transitions: its local excess is -0.019 $[-0.027, -0.008]$. Predictive likelihood improves at the boundary, with an unresolved accuracy change.

Table A15: Paired context-length comparison in Evo 2 7B. The same readout positions and 24 kb overlap groups are used at both lengths. Functional values are differences with paired 95% group intervals; negative NLL changes indicate improvement. Excess subtracts the mean of the preceding three adjacent changes.

Measurement	6 kb	24 kb
Transition norm ratio	230.8	247.1
Rank, b27	478.70	509.48
Rank, b28	11.07	11.15
Rank, b30	1.89	1.87
Rank, final embedding	3.29	3.21
Annotation AUROC, b27→b28	-0.029 $[-0.040, -0.018]$	-0.041 $[-0.054, -0.029]$
Annotation excess change	-0.031 $[-0.042, -0.018]$	-0.042 $[-0.054, -0.028]$
Next-base accuracy, b27→b28	$+0.017$ $[+0.008, +0.028]$	$+0.006$ $[-0.002, +0.016]$
Accuracy excess change	-0.007 $[-0.016, +0.006]$	-0.019 $[-0.027, -0.008]$
Next-base NLL, b27→b28	-0.053 $[-0.067, -0.038]$	-0.041 $[-0.053, -0.029]$
NLL excess change	-0.001 $[-0.018, +0.015]$	$+0.010$ $[-0.004, +0.022]$
Annotation AUROC, b28→final	-0.081 $[-0.099, -0.066]$	-0.086 $[-0.109, -0.067]$
Next-base accuracy, b28→final	$+0.013$ $[+0.004, +0.020]$	$+0.024$ $[+0.016, +0.032]$
Next-base NLL, b28→final	-0.059 $[-0.067, -0.051]$	-0.062 $[-0.070, -0.053]$

The 24 kb-minus-6 kb difference in excess annotation change is -0.011 $[-0.018, -0.001]$; the corresponding excess-accuracy difference is -0.012 $[-0.022, -0.002]$. These paired contrasts characterise the supplied-context extension in 7B on this human panel. The extension retains concentration and annotation steepening; transition accuracy is unresolved, and final output-head alignment was not remeasured. Species transfer, megabase contexts and the corresponding 40B extension remain outside this evaluation.

C Annotation within similar predictive contexts

This appendix supports Section 3.1. The analysis compares biological readout among positions with similar intact next-base predictions. It reuses the 32 disjoint chr22 windows of the dimension-restriction confirmation panel (Appendix D), called panel D in the experiment records. The union contains 12,259 unique positions from four annotation-versus-intron tasks. The pre-transition, transition and minimum-rank states are b27/b28/b30 in 7B and b20/b21/b23 in 40B; final denotes the separate RMSNorm output. Each state uses its existing no-PCA, $C = 1$ logistic out-of-fold annotation scores. One intact four-base probability vector per position defines the comparisons at every state. The analysis plan was fixed before this conditioning analysis; the panel's earlier dimension-restriction results were already available.

Stratification and matching. The first estimator forms 16 strata from the most probable nucleotide and quartiles of predictive entropy. Quartile boundaries use the union rows. Within each task, strata with fewer than 20 positives or 20 negatives are excluded; AUROCs are combined with weights equal to the number of positive–negative pairs. The second estimator greedily matches annotation positives to negatives without replacement, in increasing order of symmetric KL distance, $D_{\mathrm{KL}}(p\|q) + D_{\mathrm{KL}}(q\|p)$. Its caliper is the 10th percentile of all positive–negative distances pooled across the four tasks in that model. Matched concordance is the proportion of retained pairs whose positive has the larger annotation score, with ties counting one half. It is distinct from the all-pairs AUROC. Strata and matched pairs remain fixed across states.

The calipers are $0.1475/0.1819$ nats for 7B/40B. Mean distances of retained pairs range from 0.0106 to 0.0160 in 7B and 0.0115 to 0.0197 in 40B. Matching retains 80.7–93.0% and 70.0–92.8% of task positives, respectively. Only the splice-donor stratum with most probable base T and highest entropy is excluded; it accounts for $4.0/3.5\%$ of donor-task rows and $1.16/1.08\%$ of within-stratum pair mass.

Pooling and conditional uncertainty. The prespecified estimators pool out-of-fold scores across the five folds. After examining those results, we added a sensitivity analysis that computes AUROC within folds and averages the five scores. Stratified sensitivity uses the same task-level stratum exclusions; matched sensitivity retains the original pairs whose members share a fold and averages concordance within folds. It does not rematch within folds. The latter uses 200–351 pairs per task, or 16.1–22.0% of original pairs in 7B and 16.4–24.8% in 40B. This comparison addresses score offsets and scales between independently fitted probes.

All contrasts use 2000 shared exponential genomic-group weights, with equal task weights. Pair contributions receive the product of their members' group weights. Intervals condition on fitted probes, strata and matching; they do not refit probes or estimate new strata or pairs. Similar predicted distributions are an observational conditioning variable, so these comparisons do not identify causal mediation by prediction.

Table A16: Evo 2 7B, panel D: annotation contrasts conditional on intact predictions. Stratified entries are AUROC changes; matched entries are concordance changes. Rows contrast the pre-transition, transition and minimum-rank states b27/b28/b30 and the final embedding. Paired 95% intervals condition on fitted probes, fixed strata and pairs.

Contrast	Unstratified	Stratified	Matched
Prespecified: pooled OOF scores			
b27 → b28	−0.029 [−0.046, −0.015]	−0.038 [−0.056, −0.022]	−0.049 [−0.075, −0.030]
b28 → b30	−0.093 [−0.135, −0.048]	−0.110 [−0.158, −0.058]	−0.107 [−0.162, −0.050]
b27 → final	−0.081 [−0.125, −0.042]	−0.105 [−0.157, −0.060]	−0.109 [−0.162, −0.063]
Post-hoc: within-fold scores			
b27 → b28	−0.027 [−0.041, −0.016]	−0.029 [−0.047, −0.016]	−0.056 [−0.103, −0.021]
b28 → b30	−0.086 [−0.124, −0.059]	−0.088 [−0.130, −0.052]	−0.098 [−0.158, −0.042]
b27 → final	−0.090 [−0.123, −0.071]	−0.100 [−0.141, −0.073]	−0.131 [−0.189, −0.091]

Table A17: Evo 2 40B, panel D: annotation contrasts conditional on intact predictions. Stratified entries are AUROC changes; matched entries are concordance changes. Rows contrast the pre-transition, transition and minimum-rank states b20/b21/b23 and the final embedding. Paired 95% intervals condition on fitted probes, fixed strata and pairs.

Contrast	Unstratified	Stratified	Matched
Prespecified: pooled OOF scores			
b20 → b21	−0.020 [−0.045, +0.003]	−0.042 [−0.075, −0.013]	−0.046 [−0.089, −0.014]
b21 → b23	−0.125 [−0.162, −0.089]	−0.158 [−0.204, −0.100]	−0.168 [−0.228, −0.091]
b20 → final	−0.127 [−0.179, −0.080]	−0.162 [−0.222, −0.098]	−0.182 [−0.256, −0.100]
Post-hoc: within-fold scores			
b20 → b21	−0.013 [−0.035, +0.004]	−0.016 [−0.044, +0.005]	−0.026 [−0.077, −0.002]
b21 → b23	−0.113 [−0.143, −0.089]	−0.125 [−0.157, −0.085]	−0.154 [−0.204, −0.072]
b20 → final	−0.113 [−0.152, −0.083]	−0.124 [−0.167, −0.074]	−0.180 [−0.241, −0.117]

Both scales retain transition-to-minimum and pre-to-final decreases under all six estimators. The 7B initial decrease also persists throughout. For 40B, the initial unstratified interval includes zero, whereas pooled stratified and matched intervals exclude zero. Within-fold stratification leaves that contrast unresolved; same-fold matching retains a negative interval. The mean decline also varies by task: 40B coding-exon readout improves across the initial transition.

A second panel with fixed inputs and size. We repeat the intact-state comparison on 32 new 6 kb chr22 windows, called panel C2 in the experiment records. Eight windows are centred on each of the four annotation classes; sampling excludes the original 100 chr22 windows, 12 basis windows, eight pilot windows and all 32 panel-D inputs. The panel's size, inputs, task rows and folds were locked before its first forward evaluation, after earlier panel-D outcomes were available. Its size was not increased after observing results. The four tasks contain 4000, 3180, 3140 and 4000 rows (coding

exon, splice donor, splice acceptor and background), with 12,165 unique union positions. Each state uses independently fitted no-PCA $C = 1$ logistic probes on five grouped folds, following the panel-D protocol.

Table A18 reports the intact trajectory. Fold-mean initial contrasts are -0.0343 $[-0.0460, -0.0207]$ in 7B and -0.0576 $[-0.0707, -0.0445]$ in 40B; the transition-to-minimum and pre-to-final decreases also have negative intervals. This replication uses new inputs; it does not change the unresolved initial 40B contrast on panel D. Panel C2 is distinct from the residual-precision experiment, which uses panel D (Appendix E.3).

Table A18: Intact-state annotation AUROC on the second 32-window chr22 panel (C2). Scores average held-out AUROC within folds and then across four tasks; paired 95% genomic-group intervals hold fitted probes fixed. The pre-transition, transition and minimum-rank states are aligned by role across models.

State	AUROC [95%]
■ Evo 2 7B	
Pre-transition (b27)	0.8417 [0.8094, 0.8731]
Transition (b28)	0.8075 [0.7774, 0.8389]
Minimum-rank (b30)	0.6966 [0.6521, 0.7316]
Final embedding	0.7008 [0.6642, 0.7301]
■ Evo 2 40B	
Pre-transition (b20)	0.8529 [0.8264, 0.8857]
Transition (b21)	0.7953 [0.7632, 0.8341]
Minimum-rank (b23)	0.7077 [0.6660, 0.7459]
Final embedding	0.6920 [0.6518, 0.7314]

Predictive-context comparison on the second panel. Entropy boundaries and matching calipers are transferred unchanged from panel D, as specified in the conditioning plan; they are not refitted to C2. Strata and pairs are formed on the new positions and held fixed across states. All three contrasts retain negative intervals under pooled and within-fold estimators in both models (Tables A19 and A20). Thus the later declines persist under prediction conditioning on both panels, while the initial 40B contrast remains panel- and estimator-dependent. The intact and conditioned results on C2 share inputs and are not separate replications.

Table A19: Evo 2 7B, panel C2: annotation contrasts conditional on intact predictions. Stratified entries are AUROC changes; matched entries are concordance changes. Rows contrast the pre-transition, transition and minimum-rank states b27/b28/b30 and the final embedding. Paired 95% intervals condition on fitted probes, fixed strata and pairs.

Contrast	Unstratified	Stratified	Matched
Prespecified: pooled OOF scores			
b27 → b28	-0.040 $[-0.063, -0.015]$	-0.053 $[-0.080, -0.023]$	-0.063 $[-0.097, -0.031]$
b28 → b30	-0.117 $[-0.152, -0.086]$	-0.147 $[-0.176, -0.118]$	-0.142 $[-0.177, -0.104]$
b27 → final	-0.151 $[-0.183, -0.118]$	-0.182 $[-0.214, -0.145]$	-0.191 $[-0.239, -0.148]$
Post-hoc: within-fold scores			
b27 → b28	-0.034 $[-0.046, -0.021]$	-0.048 $[-0.063, -0.030]$	-0.044 $[-0.100, -0.013]$
b28 → b30	-0.111 $[-0.142, -0.092]$	-0.123 $[-0.152, -0.101]$	-0.105 $[-0.160, -0.069]$
b27 → final	-0.141 $[-0.169, -0.119]$	-0.162 $[-0.191, -0.134]$	-0.155 $[-0.215, -0.116]$

Table A20: Evo 2 40B, panel C2: annotation contrasts conditional on intact predictions. Stratified entries are AUROC changes; matched entries are concordance changes. Rows contrast the pre-transition, transition and minimum-rank states b20/b21/b23 and the final embedding. Paired 95% intervals condition on fitted probes, fixed strata and pairs.

Contrast	Unstratified	Stratified	Matched
Prespecified: pooled OOF scores			
b20 → b21	-0.073 $[-0.097, -0.049]$	-0.112 $[-0.137, -0.084]$	-0.124 $[-0.151, -0.091]$
b21 → b23	-0.100 $[-0.136, -0.061]$	-0.144 $[-0.181, -0.103]$	-0.157 $[-0.201, -0.109]$
b20 → final	-0.187 $[-0.231, -0.145]$	-0.270 $[-0.308, -0.225]$	-0.293 $[-0.341, -0.236]$
Post-hoc: within-fold scores			
b20 → b21	-0.058 $[-0.071, -0.045]$	-0.089 $[-0.103, -0.068]$	-0.104 $[-0.134, -0.059]$
b21 → b23	-0.088 $[-0.114, -0.071]$	-0.109 $[-0.135, -0.081]$	-0.126 $[-0.184, -0.087]$
b20 → final	-0.161 $[-0.193, -0.137]$	-0.230 $[-0.262, -0.200]$	-0.228 $[-0.279, -0.173]$

D DIMENSION RESTRICTION AND NATIVE FUNCTION

This appendix supports Section 3.2: the restriction protocol with its predefined retention margins, then the results with random-basis controls.

D.1 PROTOCOL AND PREDEFINED MARGINS

We compare internal states immediately before transition, after transition and at minimum effective rank: b27/b28/b30 in 7B and b20/b21/b23 in 40B. Development data comprise 2520 eligible positions in 12 chr22 windows. The cached 7B probe analysis excludes every input window overlapping these development inputs. A separate eight-window native pilot precedes confirmation on 32 newly selected, mutually disjoint 6 kb windows. Confirmation inputs overlap none of the 100 original chr22 input windows. Window selection is stratified by annotation, without using model outputs. This defines an annotation-enriched evaluation population. The confirmation panel contains 12,259 unique positions. The coding-exon, donor, acceptor and background tasks contain 4000, 3200, 3220 and 4000 rows, respectively, spanning 25, 22, 22 and 29 window groups. A position can participate in more than one binary task; the task-average native metric uses the same repeated contributions.

State intervention. Development positions determine each state's mean μ, coordinate standard deviations D and principal directions Q_k after coordinate standardisation. At every nucleotide position in an evaluation input, we replace the selected block output by

$$\widetilde{h}_t^{(k)} = \mu + DQ_kQ_k^\top D^{-1}(h_t - \mu), \qquad k \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}. \tag{5}$$

Here Q_k contains the leading k eigenvectors of that state's standardised development covariance. The directions are fitted separately for each model and state and shared across evaluation positions. Thus $k = 16$ denotes a 16-dimensional subspace of position-varying states, not 16 selected coordinates or an effective-rank estimate. The development mean and original hidden width are preserved; downstream computation can change the dimension again. Three Gaussian orthonormal random bases provide controls at $k = 4, 16, 64$; these match dimension and have different retained energy and perturbation size. Comparisons across states likewise match k, not retained variance or edit magnitude. The confirmation protocol was fixed after the cached curves and first two native pilot states, retaining the complete previously specified dimension grid; the same states and conditions were replicated at 40B. The $k = 16$ result summarises a point on that grid, rather than an estimated minimum sufficient dimension.

Each edited state is cast to the block's native output dtype. All subsequent blocks, final normalisation and the pretrained head are recomputed. Only the unmodified prefix is reused; no downstream activation or inference cache is reused. An unedited continuation must reproduce all four-base logits at every position of every evaluated window exactly. The 40B runtime restores FP8 projection extra state before each pass. The fresh 7B development activations also match the cached development activations exactly.

Paired functional measurements. Four binary annotation tasks use fixed grouped folds and matched next-base targets. Native four-base NLL and forward $\mathrm{KL}(p_{\mathrm{intact}} \| p_{\mathrm{edited}})$ are evaluated on the same rows as annotation. Separate linear probes are refitted for every condition at the local edited state and final state. Training-fold coordinate standardisation precedes $C = 1$ logistic regression; no additional PCA discards edited feature directions. The cached analysis uses an equivalent orthonormal compact representation, verified to preserve feature inner products, the ℓ_2 penalty and fitted probabilities. Forward analyses fit directly to the actual tensors after casting. Scores average held-out folds and then tasks equally. Native summaries over unique positions are retained separately in the measurement files.

Paired intervals use 2000 exponential genomic-group weights shared across states and conditions, conditional on fitted probes and bases. Before pilot outcomes, we set native retention criteria to a 95% NLL-change interval contained in $[-0.01, 0.01]$ nats and an upper 95% bound on mean forward KL of at most 0.01 nats. Annotation retention uses a 95% AUROC-change interval contained in $[-0.01, 0.01]$. We report continuous effects throughout the grid. These pointwise criteria do not support searching the same grid for a minimum sufficient dimension. Local and final geometry use the same fixed 1024-row subset; per-position realised perturbation norms and output-state norms are retained for every arm.

D.2 Results and random-basis controls

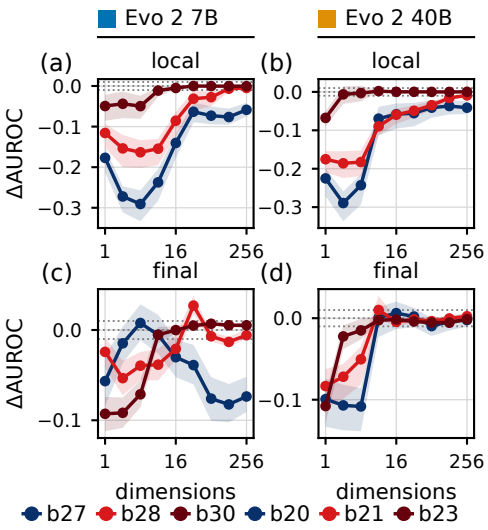

Figure A2: Paired annotation changes under state dimension restriction: (a, b) local state and (c, d) final state, one column per model. Each curve compares its restricted state with its own intact reference (edited minus intact), complementing the four-task mean AUROCs in Figure 4. Bands are pointwise paired 95% genomic-group intervals conditional on fitted probes and bases; the dotted guides at ± 0.01 mark the predefined annotation-retention margin.

Panel-specific intact baselines. On the new confirmation panel, intact transition-minus-pre-transition annotation changes are -0.0272 $[-0.0412, -0.0162]$ in 7B and -0.0133 $[-0.0346, 0.0041]$ in 40B. The latter interval includes zero, so this panel does not independently resolve the primary 40B transition decline of -0.036 $[-0.047, -0.027]$. Subsequent changes to the minimum-rank states are -0.0865 $[-0.1239, -0.0594]$ and -0.1133 $[-0.1435, -0.0893]$. These are separate estimates from the larger original trajectory panels. The dimension intervention compares each restricted representation with its own intact state. At the transition, the 16-direction annotation losses are -0.0856 $[-0.1216, -0.0647]$ and -0.0594 $[-0.0828, -0.0436]$; both models meet the native NLL and KL criteria. At 40B b21, the NLL-change interval is $[-0.000249, 0.000315]$ nats and the KL interval is $[0.000162, 0.000213]$ nats. Native retention therefore follows from explicit margins, rather than a failure to detect a difference. The absolute intact annotation levels also decline across stages, so retention relative to a state's baseline does not imply equal biological utility across states.

Stage-dependent dimension effects. Table A21 reports the $k = 16$ comparison for all six states, including the native effects, local annotation changes and paired intervals. The pre-transition restrictions cost 0.102 nats in 7B and 0.065 nats in 40B and fail native retention; at the minimum-rank

states b30/b23, the same restriction meets both native and local annotation criteria. With only four directions at 40B b23, however, NLL increases by 0.343 $[0.297, 0.385]$ nats and output KL is 0.324 $[0.282, 0.361]$ nats. Thus a low entropy effective rank does not specify how many principal directions the remaining native computation needs (Figure 4).

Table A21: Native prediction and local annotation at $k = 16$ in the predefined dimension grid. Each cell gives an estimate and its pointwise 95% genomic-group interval, conditional on fitted probes and bases. Changes subtract the corresponding intact state; KL is $\mathrm{KL}(p_{\mathrm{intact}} \| p_{\mathrm{edited}})$. Native columns are in 10^{-3} nats. Scores average four tasks and five folds equally on identical rows. Native retention requires the NLL-change interval within $[-10, 10]$ and the KL upper bound at most 10 in these units; annotation retention requires the AUROC-change interval within $[-0.01, 0.01]$. Full grids, final-state readouts and controls accompany the numerical artifact.

Model	State	ΔNLL (10^{-3} nats)	KL (10^{-3} nats)	ΔAUROC
■ 7B	b27	102.0 $[79.6, 121.6]$	101.8 $[81.9, 118.9]$	-0.1404 $[-0.1783, -0.1188]$
■ 7B	b28	6.269 $[4.523, 7.901]$	5.177 $[4.364, 5.763]$	-0.0856 $[-0.1216, -0.0647]$
■ 7B	b30	0.005 $[-0.014, 0.024]$	0.001 $[0.001, 0.001]$	-0.0049 $[-0.0080, -0.0018]$
■ 40B	b20	65.1 $[48.9, 83.4]$	65.2 $[51.7, 79.3]$	-0.0582 $[-0.0952, -0.0301]$
■ 40B	b21	0.021 $[-0.249, 0.315]$	0.188 $[0.162, 0.213]$	-0.0594 $[-0.0828, -0.0436]$
■ 40B	b23	-0.003 $[-0.018, 0.009]$	0.001 $[0.000, 0.001]$	0.0002 $[-0.0001, 0.0005]$

Figures A3 and A4 compare PCA restriction with three random subspaces at matched retained dimensions in 7B and 40B. Matched dimensions do not imply matched perturbation sizes.

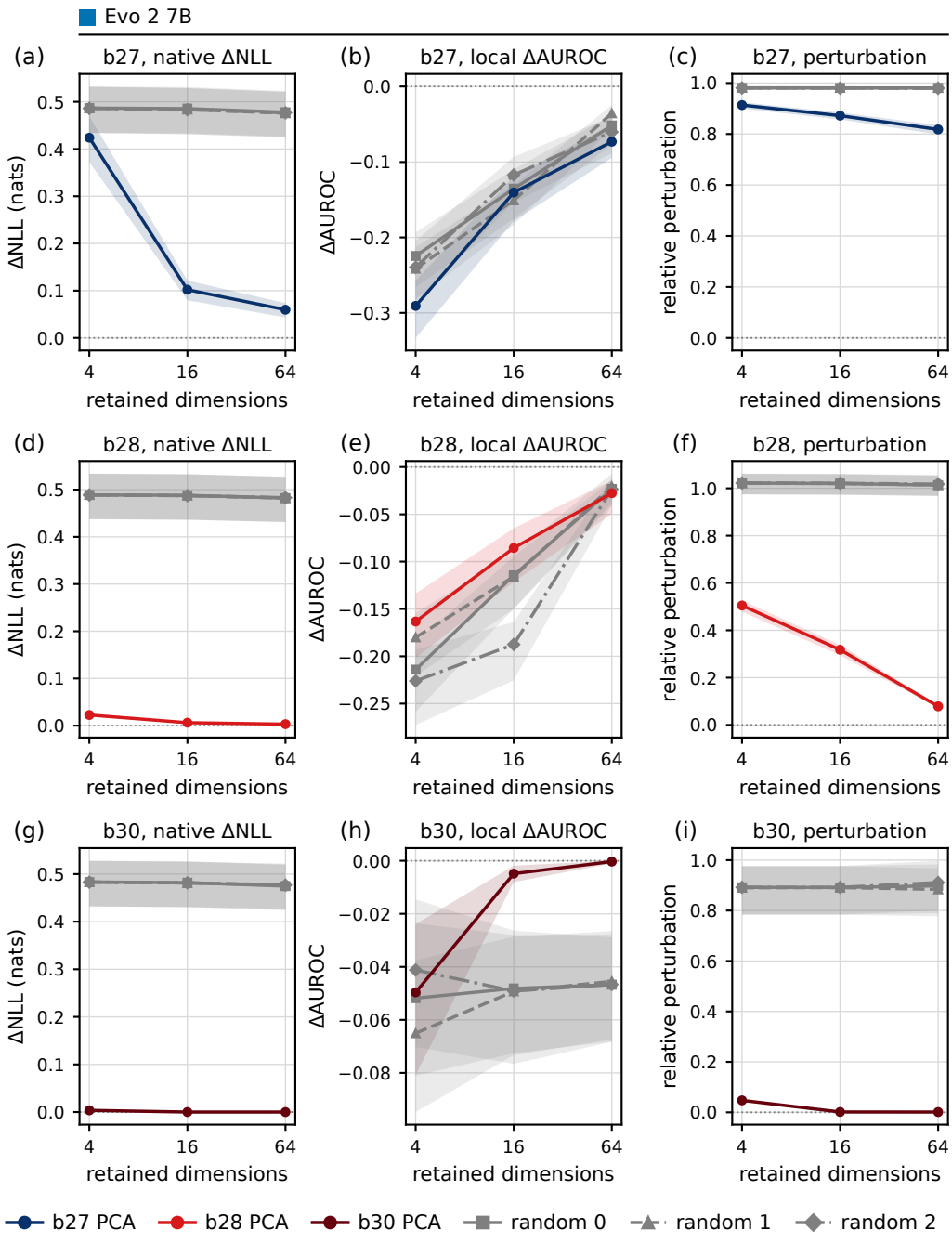

Figure A3: 7B dimension controls. Rows are states b27, b28 and b30; columns show native NLL change, local AUROC change (both edited minus intact), and relative perturbation. Random subspaces are grey and differ by line style. PCA and three independent random subspaces have the same dimension and different perturbation sizes. NLL and annotation changes share the same rows; relative perturbation is the per-position edit norm divided by the intact output-state norm, then averaged with the task/fold weights. Bands are pointwise paired 95% intervals.

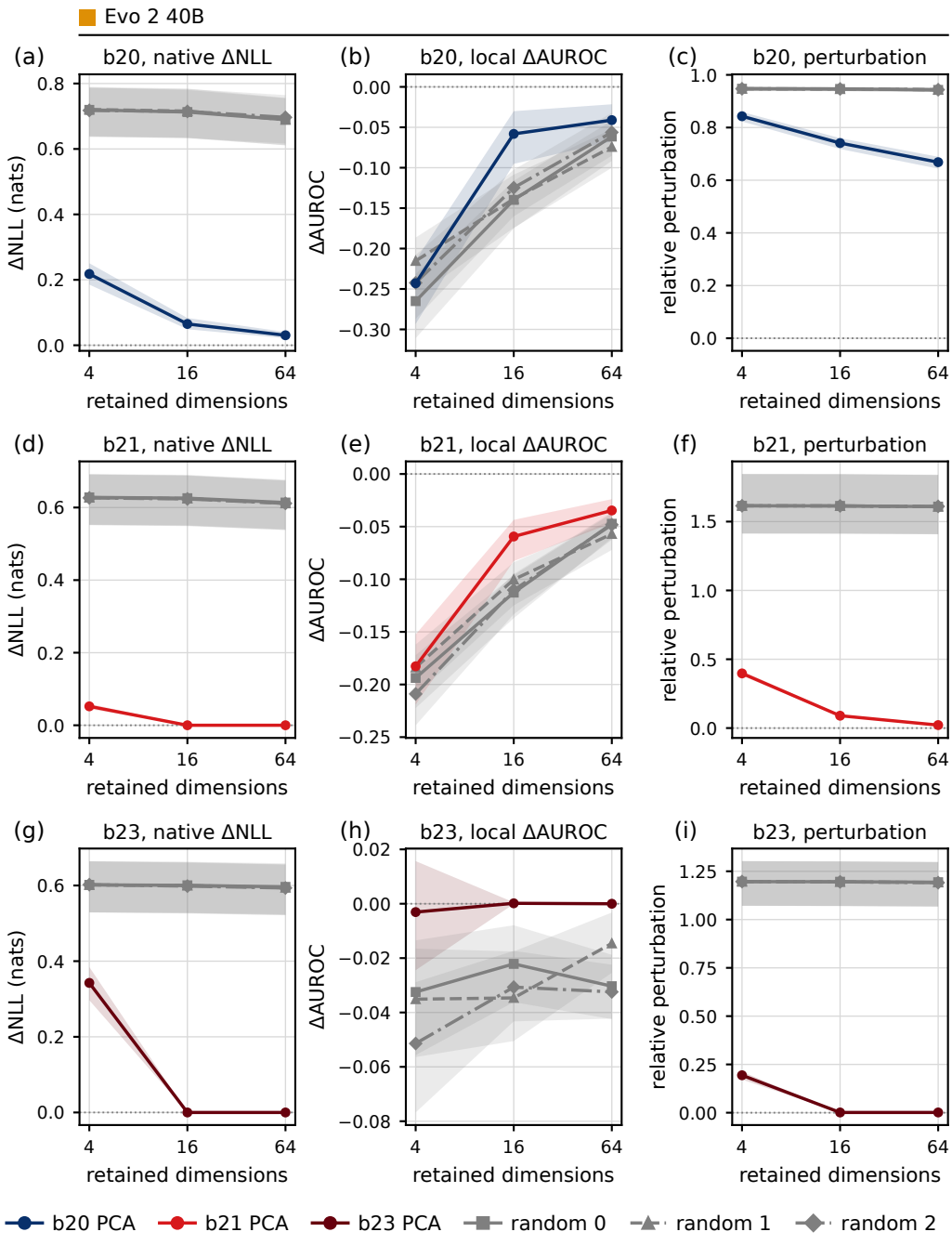

Figure A4: 40B dimension controls at b20, b21 and b23, with the same definitions as Figure A3.

E STAGED GEOMETRY

This appendix supports Section 3.3: the fixed-row profiles separate expansion, minimum rank and output alignment, and precision controls test the role of numerical rounding.

E.1 CANONICAL GEOMETRY, LATE INERT BLOCKS AND ROUNDING

The canonical table is recomputed from fixed rows and one metric implementation. It separates raw update dimension, centred update dimension, standardised state dimension and output alignment, whose A column is the centred ACGT logit-contrast fraction of rank 3. The first expansion is not the largest expansion in Evo 2, and a realised update that rounds to zero is reported as zero rather than omitted from a logarithmic trajectory. Final normalisation is a separate endpoint.

Table A22: Canonical fixed-row metrics. The three rank columns refer to different matrices: the realised update Δ, its position-centred version Δ_c, and the coordinate-standardised state $Z(H)$. A is centred rank-3 head-contrast alignment. Exact zero updates are written as 0; small nonzero updates use scientific notation. Endpoint transitions include final normalisation.

				Effective rank			
Model	State	ρ	ω	Δ	Δ_c	$Z(H)$	A (%)
Evo 2 7B	b27	1.48	0.96	279.23	392.99	478.70	0.09
Evo 2 7B	b28	230.80	230.65	4.69	8.71	11.07	0.59
Evo 2 7B	b29	799.83	799.50	2.86	4.64	4.90	0.14
Evo 2 7B	b30	6.20×10^{5}	6.20×10^{5}	1.19	1.66	1.89	10.58
Evo 2 7B	b31	1.00	0	0	0	1.89	10.58
Evo 2 7B	norm	5.73×10^{-12}	1.00	1.19	1.66	3.29	63.21
Evo 2 40B	b20	2.25	1.93	82.47	127.17	329.95	0.15
Evo 2 40B	b21	295.24	295.22	3.93	5.86	6.77	5.47
Evo 2 40B	b22	1.26	0.46	1.51	4.22	7.38	5.24
Evo 2 40B	b23	3.09×10^{13}	3.09×10^{13}	2.27	2.97	3.89	3.86
Evo 2 40B	b24	1.00	0	0	0	3.89	3.86
Evo 2 40B	b34	1.03	0.24	1.00	1.00	4.08	3.77
Evo 2 40B	b49	1.00	1.30×10^{-22}	99.76	99.81	4.08	3.77
Evo 2 40B	norm	2.27×10^{-18}	1.00	2.28	3.01	5.56	34.43
CENO 1B	b35	1.22	0.83	15.61	20.33	167.02	2.18
CENO 1B	b36	9018.47	9018.17	1.16	1.18	2.19	0.48
CENO 1B	b37	1.00	1.42×10^{-3}	3.67	4.32	2.19	0.48
CENO 1B	norm (b38)	5.03×10^{-5}	1.00	1.16	1.18	43.49	24.39

A zero realised update denotes an exactly zero stored state difference on the sampled rows. A nonzero branch tensor can produce this result through rounding during residual addition. This occurs at Evo 2 7B b31 and 40B b24 in Table A22, where zero matrices are assigned effective rank zero. At 7B b31, the separate arithmetic control confirms that nonzero branch writes are absorbed by addition to the much larger residual (Appendix E.3). The check covers sampled states; full-tensor bit identity was not evaluated.

Small nonzero updates are written in scientific notation. CENO b37 has $\omega = 1.42 \times 10^{-3}$ and nonzero raw/centred update ranks; 40B b49 has $\omega = 1.30 \times 10^{-22}$. CENO b36 and b37 both have standardised state rank 2.19 to the reported precision, yet their annotation point estimates differ in all six tasks (Table A29). Equal effective rank therefore does not identify equal states.

Vocabulary-suppression direction. The normalised difference between the mean A/C/G/T head row and the mean remaining vocabulary row tests one possible source of concentration. Transition rank remains low after projecting this direction out of the hidden state, so this direction alone does not explain the concentration; the test does not identify its mechanism.

E.2 INPUT COMPOSITION

For a single real chr22 window and its random, poly-A and shuffled replacements, the 7B b28 gated-MLP norm exceeds the sum of incoming-state and mixer mean norms by 61.5, 48.6, 112.8 and 43.0 times, respectively. The expansion therefore persists under these substitutions, but this one-window control measures no annotation readout.

E.3 ARITHMETIC PRECISION CONTROLS

Three controls change only residual or projection arithmetic while retaining native branch computation. The first two measure geometry and native prediction on 12 fixed windows; the third adds annotation readout on identical positions.

7B residual precision and local rounding. The late rank decreases persist when residual accumulation is performed in FP32. The controls distinguish loss of small writes during residual addition from changes to branch computation or the pretrained runtime. The tested modes retain native branch dtypes.

The precision panel uses 12 fixed chr22 windows and every 24th nucleotide, with the same 2048 row subset for all state spectra. Four full forward modes differ only in the indicated arithmetic: native; FP32 residual sums and retained stream; FP32 sums with block-output casts back to bf16; and native residuals with a separate FP32 final-normalisation calculation. Branch normalisation inputs are cast to the dtype captured in the native run; branch matrix multiplications retain native precision. Rewritten addition sites in the provider runtime are checked by an instrumented native no-op.

Table A23: Full forward precision comparison on identical 12-window inputs. Only the named residual/norm operations change; branch computation remains native. Head NLL (nats) and accuracy use the pretrained head, not fitted probes.

Mode	Rank b28	Rank b30	Rank norm	Head NLL	Head acc.
Native	8.4118	1.8948	5.0654	1.035126	0.549731
FP32 residual stream	8.4059	1.8945	5.0666	1.035163	0.549536
FP32 sums, block cast-back	8.4096	1.8948	5.0655	1.035141	0.549453
FP32 final norm only	8.4118	1.8948	4.9353	1.035119	0.549717

Table A23 retains the numerical differences between modes: native and FP32-stream standardised rank at block 30 differ in the fourth decimal; the FP32 stream increases next-token NLL by 3.73×10^{-5} nats.

For each sampled addition at blocks 26–31, we record the intended branch tensor, realised $(x+d)-x$, unchanged-coordinate fraction and the fraction of intended entries below half a ULP of the residual. Relative rounding error is $\|((x+d)-x)-d\|/\|d\|$, evaluated in FP64; this expression avoids subtracting two large sums before detecting a small lost write. The local FP64 replay recomputes each cached addition with its recorded operands, keeping branch outputs fixed and omitting downstream execution. At b31, all sampled coordinates remain unchanged in native mode; at least 99.9999% remain unchanged in each FP32-stream addition. The nonzero branch tensors are thus almost entirely absorbed by residual rounding (Appendix E.1). The mixer changes most coordinates in FP64 replay; the MLP is too small even for this replay. Fp32 residual accumulation therefore retains substantial rounding of small branch updates. Model accuracy and NLL are nearly unchanged across the tested modes; annotation readout under these arithmetic modes remains unmeasured.

40B projection and residual-stream precision. On 12 fixed chr22 windows of 6000 bases, we test whether residual-stream or projection precision changes the late rank decreases. Sampling uses every 24th position and 2048 common rows for geometry. These uniformly spaced rows differ from the annotation-enriched canonical panel.

Disabling FP8 changes both the released checkpoint's execution path and its predictive behaviour. We measure prediction and geometry together under this change. The three modes separate two numerical choices. The native mode retains FP8 input projections and a bf16 residual stream. The

second mode disables FP8 input projections in all 42 affected `TELinear` modules, preserving the same weights, native bf16 stream and provider FP32 internals. The third mode retains native FP8 projections and casts residual additions and the stream to FP32; inputs to branch normalisations and final normalisation are cast to their native dtypes. Checkpoint FP8 metadata are restored before every window and mode. A no-op wrapper reproduces every captured state and output logit exactly. Observed block outputs are bf16 in the first two modes and FP32 in the third; final-normalisation output is bf16 in all modes. These modes isolate projection and residual arithmetic while retaining native branch dtypes.

Table A24: Evo 2 40B arithmetic controls on the fixed 12-window panel. R_ℓ is standardised effective rank at block ℓ, and R_N is rank after final normalisation. Native-head accuracy and NLL use all 71988 eligible adjacent-base pairs and the unchanged full-vocabulary prediction head; they are distinct from the fitted next-base probe.

Arithmetic	R_{20}	R_{21}	R_{23}	R_N	Accuracy	NLL
Native	223.37	6.23	3.30	4.83	0.623	0.906
FP8 input projections disabled	58.36	6.10	3.01	3.05	0.327	1.359
FP32 residual stream	223.63	6.22	3.30	4.83	0.623	0.906

The FP32 residual stream closely reproduces native geometry and prediction performance. Disabling FP8 input projections also retains low rank after b21, but materially reduces native-head accuracy and increases NLL (Table A24). The corresponding adjacent norm ratios at b21 are 235.3, 59.3 and 235.0 in table order; the later b23 ratios are 4.485×10^{13}, 2.103×10^{14} and 4.489×10^{13}. Low-rank geometry therefore persists across these arithmetic modes despite their different predictive behaviour; overall precision independence remains unestablished.

The uniformly spaced position panel therefore retains the b21 expansion ($\rho_{21} = 235.3$ in native arithmetic), despite sampling different positions from the annotation-enriched canonical panel.

Precision and annotation readout on identical positions. We extend the 12-window arithmetic analysis to all 32 panel-D inputs, measuring biological readout and native prediction together. Three modes use the same frozen 40B checkpoint: native FP8 projections with a bf16 residual stream; FP32 residual accumulation with native projections; and bf16 projections with a native stream. The second mode retains FP32 between residual additions but casts inputs to branch normalisation and final RMSNorm to their native dtype. It is a residual-stream control, not an all-FP32 model.

A native-mode implementation of the residual-addition control exactly reproduces the unmodified forward outputs on the first window, validating the intervention code.

States b20, b21, b23 and the final embedding are recorded at all 12,259 union rows. Each task/state/mode refits training-fold standardisation and $C = 1$ logistic probes without PCA. Annotation and probe-next-base metrics average folds and then tasks; native metrics on these rows use the same task/fold weighting. Separately, native endpoints are averaged over all 175,584 eligible positions after each window's first 512 positions. Both analyses use 2000 paired genomic-group weights. The task-weighted native baseline NLL is 0.86667; the unweighted unique-union mean is 0.85419. These are different estimands on the same forward outputs. Table A25 uses the all-position population.

Table A25: Native prediction on 175,584 eligible positions after the first 512 positions of each of 32 windows. NLL and intact-to-mode KL are in nats; intervals resample input-window groups.

Mode	NLL	ΔNLL [95%]	Accuracy	KL
Native	0.92412	+0.00000 [+0.00000, +0.00000]	0.59454	0.00000
FP32 stream	0.92413	+0.00001 [−0.00006, +0.00008]	0.59422	0.00014
bf16 projections	1.44674	+0.52262 [+0.44903, +0.61222]	0.24557	0.50128

Fp32 residual accumulation changes all-position NLL by $+0.000008$ $[-0.000059, +0.000082]$ nats, with mean output KL 0.000139 $[0.000093, 0.000214]$. Accuracy changes by -0.000319

Table A26: Four-task annotation AUROC under three 40B arithmetic modes. Probes are refitted at each state and mode; the FP32-stream mode retains native projections.

Mode	b20	b21	b23	Final
Native	0.7894	0.7761	0.6627	0.6759
FP32 stream	0.7917	0.7760	0.6627	0.6716
bf16 projections	0.5526	0.5685	0.5227	0.5101

Table A27: Paired annotation contrasts in native and FP32-residual modes. The last column is a difference of depth contrasts, with its paired 95% interval.

Contrast	Native	FP32 stream	FP32 − native
b20 → b21	-0.0133 $[-0.0346, +0.0041]$	-0.0157 $[-0.0363, +0.0020]$	-0.0024 $[-0.0033, -0.0013]$
b21 → b23	-0.1133 $[-0.1435, -0.0893]$	-0.1133 $[-0.1436, -0.0896]$	-0.0000 $[-0.0009, +0.0013]$
b20 → final	-0.1135 $[-0.1524, -0.0834]$	-0.1201 $[-0.1587, -0.0904]$	-0.0066 $[-0.0115, -0.0019]$
b21 → final	-0.1002 $[-0.1275, -0.0783]$	-0.1044 $[-0.1315, -0.0826]$	-0.0043 $[-0.0093, +0.0001]$

$[-0.000573, -0.000051]$. The transition-to-minimum and pre-to-final annotation decreases remain clear, while the initial b20-to-b21 interval includes zero under both modes. The mode difference in the initial annotation contrast is -0.0024 $[-0.0033, -0.0013]$, and in the pre-to-final contrast -0.0066 $[-0.0115, -0.0019]$. The broad pattern persists with small, measurable mode effects. These annotation differences do not bound every metric: the initial next-base probe-NLL contrast differs by -0.01115 $[-0.01471, -0.00681]$ nats.

Switching projections to bf16 substantially changes the native predictor, increasing all-position NLL by 0.5226 nats. Its low annotation scores therefore describe an altered execution mode rather than a high-precision reference for the native model.

F CENO RECOVERY AND MODEL SCOPE

This appendix supports the CENO comparison in Section 3.3 and the model scope stated in Section 5. Evo 2 7B and 40B provide the principal scale comparison. CENO tests whether annotation readout can recover while effective rank remains low; HyenaDNA provides a smaller model with a related sequence operator. These models were selected after inspecting diagnostic results, and their training data, objectives and runtimes differ.

Table A28: Focused comparison on canonical chr22 rows, using the same 2048 fixed rows at every state within each model. Blocks are physical decoder blocks. Maximum adjacent norm ratios compare consecutive block outputs, excluding embedding-to-first-block and final-normalisation steps. $R_{\mathtt{norm}}$ and $A_{\mathtt{norm}}$ are final standardised rank and centred rank-3 ACGT-contrast alignment.

Model	Blocks / width	Max. ρ	$R_{\mathtt{norm}}$	$A_{\mathtt{norm}}$ (%)
■ Evo 2 7B	32 / 4096	6.20×10^{5}	3.29	63.21
■ Evo 2 40B	50 / 8192	3.09×10^{13}	5.56	34.43
■ CENO 1B	38 / 1024	9018.47	43.49	24.39
HyenaDNA medium	8 / 256	2.98	38.43	33.99

Profiles use the state conventions of Appendix A.2 and the fixed-row geometry of Appendix A.5. Rows are shared across states within each model, not across checkpoints.

F.1 CENO RECOVERY AND FINAL-STAGE DECOMPOSITION

Focal b35–b38 annotation fits use connected-window groups and the same training-fold protocol as the primary readout. Five of the six binary annotation tasks have positive final-minus-transition intervals. For 3′UTR, the estimated change is negative and its interval $[-0.050, +0.010]$ includes

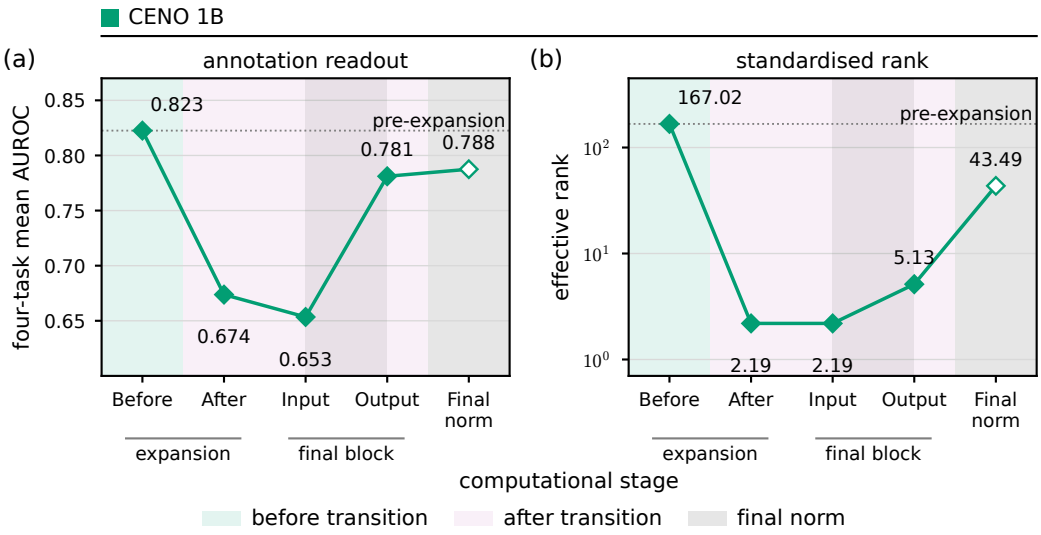

Figure A5: CENO's partial annotation recovery precedes the larger effective-rank increase at final normalisation. (a) Four-task mean annotation AUROC. (b) Standardised effective rank. Stages are cached b35, b36 and b37, the final-block output X, and final RMSNorm (b38); categorical spacing does not encode layer distance. Tints mark stages before and after expansion and at final normalisation; the grey window spans the final block's input and output, the open diamond marks final RMSNorm, and the dotted line the pre-expansion value. Annotation and geometry use distinct row samples from the same chr22 windows.

zero (Table A29). The four-task mean remains below its immediate pre-transition value (Table A6). Fitted next-base accuracy instead rises from 0.481 before transition to 0.572 at the final state, after falling to 0.308 at transition (Table A4).

Table A29: CENO annotation AUROC at focal states, using connected-window folds. Recovery intervals are paired 95% intervals for `norm` (cached b38) minus transition b36, conditional on fitted predictions.

Task vs. intron	b35	b36	b37	`norm`	Recovery 95% interval
coding exon	0.780	0.649	0.684	0.763	$[+0.081, +0.152]$
splice donor	0.930	0.768	0.735	0.906	$[+0.108, +0.165]$
splice acceptor	0.918	0.746	0.716	0.884	$[+0.090, +0.172]$
intergenic	0.662	0.532	0.479	0.597	$[+0.032, +0.108]$
5′UTR	0.891	0.676	0.698	0.892	$[+0.149, +0.248]$
3′UTR	0.790	0.668	0.619	0.652	$[-0.050, +0.010]$

Separating the final block from final normalisation. Figure A5 separates the final-block recovery from the subsequent rank increase. Here we further distinguish positionwise RMS scaling from learned coordinate gains (Table A30). The cached CENO state b38 combines the final block and final normalisation. We separate these operations by replaying the same 100 chr22 input windows in the original FP32 provider runtime. The replay retains all 47217 cached positions in their original order. The final-normalisation hook reproduces the cached b38 features bit for bit at every position. We record the final-block output X before normalisation and form four variants: X, position-dependent RMS scaling $\mathcal{S}(X)$, learned coordinate scaling XW, and the full normalisation $\mathcal{S}(X)W$, where $W = \mathrm{diag}(w)$ and

$$\mathcal{S}(X)_{ij} = \frac{X_{ij}}{\sqrt{d^{-1}\sum_{k=1}^{d} X_{ik}^2 + \epsilon}}, \qquad d = 1024, \quad \epsilon = 10^{-5}.$$

This decomposition reproduces the provider's normalisation output exactly.

Each variant is evaluated with the original assay-specific rows and connected-window folds. Annotation probes use the four binary annotation tasks against intron; the next-base probe uses its original 8000 eligible rows. Preprocessing is fitted within each training fold: StandardScaler, PCA(256), and logistic regression ($C = 1$, maximum 20000 iterations). All fits converge. The newly fitted full-normalisation predictions match the original b38 predictions for all five assays. Geometry uses the original canonical subset of 2048 rows, and reproduces the original final rank and contrast alignment.

Table A30: CENO final-block and normalisation decomposition on the original chr22 panels. Geometry uses the same canonical rows for every state; the two readout assays retain their distinct row samples. R is standardised effective rank and A is centred ACGT-contrast alignment. Annotation AUROC averages the four binary annotation tasks.

State	R	A (%)	AUROC	Probe acc.	Probe NLL
b37: final-block input	2.19	0.48	0.653	0.305	1.363
Final-block output X	5.13	1.50	0.781	0.560	1.026
RMS scaling $\mathcal{S}(X)$	43.49	22.68	0.787	0.572	0.997
Coordinate scaling XW	5.13	0.85	0.782	0.561	1.027
Full normalisation $\mathcal{S}(X)W$	43.49	24.39	0.788	0.572	0.999

The final block increases mean annotation AUROC by 0.128 (95% paired interval $[0.113, 0.146]$), next-base accuracy by 0.255 ($[0.226, 0.282]$), and decreases next-base NLL by 0.337 ($[-0.385, -0.288]$ for the signed NLL change). Its output remains low rank ($R = 5.13$). Final normalisation then increases rank to 43.49 and next-base accuracy by 0.012 ($[0.005, 0.019]$). Its annotation change is $+0.006$ ($[-0.001, 0.011]$), and its NLL change is -0.027 ($[-0.039, -0.012]$). These intervals use 2000 shared Exponential(1) weights over connected-window groups and are conditional on fitted predictions.

Position-dependent RMS scaling alone produces the rank increase and an accuracy gain of 0.012 ($[0.005, 0.019]$), with an NLL change of -0.029 ($[-0.042, -0.016]$). Adding the learned gains after RMS scaling changes accuracy by -0.00002 ($[-0.00379, 0.00366]$) and mean annotation AUROC by 0.00068 ($[-0.00056, 0.00214]$). Thus the large readout recovery occurs in the final block, before the larger rank increase produced by RMS scaling.

Feature-wise standardisation removes the magnitude of nonzero coordinate gains, leaving only sign flips; coordinate scaling alone therefore preserves the standardised spectrum in exact arithmetic. It can still change alignment with the fixed head: the learned gains increase alignment from 22.68% after RMS scaling to 24.39% after full normalisation.

F.2 HyenaDNA

HyenaDNA has a maximum adjacent block-output norm ratio of 2.98, final centred contrast alignment of 33.99%, and final standardised rank of 38.43. Thus substantial head alignment can occur without the large norm increases observed in Evo 2; differences in training and architecture prevent attribution to the sequence operator alone.

F.3 Bidirectional screens: Caduceus-PS and NT-v2

Two exploratory geometry screens extend the architectural comparison (Table A31). Caduceus-PS is a bidirectional, reverse-complement equivariant state-space model (Schiff et al., 2024); NT-v2 is a masked Transformer with predominantly six-base tokens (Dalla-Torre et al., 2025). Their maximum adjacent norm ratios are 3.78 and 3.11 on the evaluated chr22 inputs. Objectives, tokenisation, architecture and training data vary jointly, precluding attribution to a single factor.

Caduceus was re-extracted on the original 100 six-kilobase chr22 windows, using 2048 fixed annotation-enriched positions. Its block API returns a mixer output and an incoming residual separately. We sum these to recover the post-add residual, accounting for the pinned provider's channel-half permutation when checking the next block's input. That connection is exact; independently recomputing final normalisation agrees with the model output to relative maximum error

Table A31: Additional geometry screens on chr22, using 2048 fixed rows within each model. Ratios compare consecutive block outputs. NT-v2 has no separately stored final-normalised state. These are not matched tests of objective, architecture or vocabulary size.

Checkpoint	State measurement	Maximum ρ
Caduceus-PS, 16 blocks	Accumulated residual	3.78
NT-v2 500M	Cached block states	3.11

2.97×10^{-7}. The maximum ratio is 3.78 (3.77 over all 47217 retained rows). Standardised effective rank nevertheless falls from 118.9 at b14 to 21.9 at b15, and is 22.9 after final normalisation. Directional concentration therefore occurs in Caduceus without a large norm increase.

The NT-v2 row uses a 2048-row geometry screen; its cache did not separate the last block from final normalisation. We compare only its norm profile. Both screens use unmasked sequence inputs. Because their bidirectional states can depend on the target base, causal next-base probes are inapplicable.

F.4 Runtime validation and comparison limits

HyenaDNA cache and runtime check. The medium checkpoint has eight blocks and uses FP32 parameters. Replaying the original extraction reproduces cached labels and window order, and every token spans one base. A native forward pass on a fixed chr22 window reproduces all cached states at sampled positions. On its 5999 adjacent-base pairs, full-vocabulary native-head accuracy is 0.5601 and NLL is 1.0462, compared with majority-base accuracy 0.4029. This validates the predictive runtime underlying the cache.

Excluded HybriDNA comparison. The HybriDNA geometry extraction and interventions use the provider's PyTorch fallback with `use_mamba_kernels=False`, confirmed by the extraction log. On the intervention panel this runtime gives native accuracy 0.321, close to the majority-base baseline of 0.309. Equivalence to the fused runtime has not been established, so we exclude these profiles from the comparative evidence; the low accuracy alone does not establish a runtime defect.

G Earlier-state reconstruction

This appendix supports Section 3.4: the linear reconstruction of the 40B pre-transition state with its matched self-reference, then its sensitivity to source dimension, training-selected regularisation and a validation-stopped nonlinear decoder.

G.1 Linear reconstruction and the self-reference

The reconstruction assay measures recoverability of the Evo 2 40B immediate pre-transition state b20 under a fixed linear decoder. We predict all 8,192 b20 coordinates from b19, b20 itself, b21, b22, b23 or the final-normalised embedding in Evo 2 40B. The analysis uses 8,000 positions per chromosome, selected from the existing caches with a fixed random sample. The selected positions are identical across source states. Five outer folds group overlap-connected input windows, giving 66 components on chr22 and 75 on chr17. Chromosomes are evaluated separately.

Within each training fold, source features are coordinate-standardised and projected to 256 principal components (randomised SVD). The target is coordinate-standardised using training rows and retains all 8,192 coordinates. A ridge map with fixed penalty $\alpha = 1$ predicts these target coordinates. For the b20 self-reference, the input still passes through PCA256 and the fitted ridge map; it is not an identity copy of the target, so unit held-out R^2 is not expected. All scalers, the source PCA and the ridge map are fitted on training rows only. Predictions are transformed back to the original target units before evaluation.

The primary statistic averages coordinate-wise R^2 uniformly across the full target state and then averages the five held-out folds. We separately report raw variance-weighted R^2, which gives greater

Table A32: Reconstructing all 8,192 coordinates of the 40B immediate pre-transition state b20. Each value averages five held-out folds grouped by overlap-connected input windows. Source standardisation, PCA256 and ridge regression, and target standardisation, use training rows only. Coordinate-mean R^2 weights target coordinates equally; variance-weighted R^2 weights their raw held-out variances.

	chr22		*chr17*	
Source	Mean R^2	Weighted R^2	Mean R^2	Weighted R^2
b19	0.440	0.510	0.450	0.518
b20 (self-reference)	0.647	0.711	0.659	0.720
b21	0.478	0.565	0.490	0.573
b22	0.450	0.539	0.461	0.547
b23	0.149	0.219	0.154	0.218
Final embedding	0.172	0.241	0.186	0.254

weight to coordinates with greater held-out variance. Both use the held-out coordinate mean as the denominator reference. The final-normalised embedding is subject to the same source PCA256 bottleneck.

On chr22, coordinate-mean R^2 is 0.440 from b19, 0.478 from b21, 0.450 from b22, 0.149 from b23 and 0.172 from the final embedding. The b21 point estimate exceeds b19 despite its much lower rank; paired uncertainty for this comparison is unavailable. The reduction at b23 also appears on chr17 and under raw-variance weighting. The b20 self-reference, using the same source PCA256 bottleneck and train-only fitting, achieves coordinate-mean $R^2 = 0.647/0.659$ and variance-weighted $R^2 = 0.711/0.720$ on chr22/chr17. Its selected target-array hashes, row identities, folds and target sums of squares match the other sources exactly. The fitting implementation and hyperparameters are unchanged. The self-reference calibrates the scores under the source bottleneck and held-out fitting protocol. It is a decoder-performance reference, not a strict upper bound for alternative sources or a denominator for the fraction of biological information preserved. Initial rank concentration precedes the further reconstruction decline at b23. Per-fold sufficient statistics reproduce all reported metrics.

G.2 Source dimension, regularisation and a nonlinear decoder

Source dimension and training-selected regularisation. We repeat reconstruction from b20, b21 and b23 with source PCA64, PCA256, PCA1024 and no source projection on the same 8000 rows and five overlap-connected outer folds. For each outer fold, three grouped splits of the training data select the ridge penalty from $\{0.01, 1, 100, 10000\}$ by mean coordinate-wise validation R^2. Source and target scalers and source PCA are refitted in each inner split. The PCA64 and PCA256 arms use the leading directions of the fitted PCA1024 decomposition (randomised SVD). The full target retains all 8192 coordinates. Table A33 reports paired genomic-group intervals from 2000 exponential-weight resamples; each resample recomputes held-out target means and sums of squares within each fold. Intervals condition on fitted decoders and selected penalties. The PCA256 self-reference here differs slightly from Figure 5(d) because its penalty is training-selected rather than fixed at $\alpha = 1$.

Increasing the retained PCA dimension from 64 to 1024 improves reconstruction from b20 and b21. Removing PCA yields lower b21 point estimates than PCA1024 on both chromosomes. Reconstruction from b23 remains approximately 0.15 across all four source settings. The paired b23-minus-b21 interval is below zero for every source dimension on both chromosomes. The additional decline at b23 therefore persists after removing the PCA256 bottleneck and selecting regularisation within the training data.

Validation-stopped nonlinear decoder. On chr22 we also fit a PCA1024–MLP256 decoder with one GELU hidden layer and all 8192 output coordinates. A grouped validation split inside each outer training fold selects the epoch count; source and target transforms are fitted on the inner training rows and then refitted on the complete outer training fold. AdamW uses learning rate 10^{-3}, weight decay 10^{-4} and batches of 512. Training stops after 60 epochs without a validation-MSE improvement of at least 10^{-5}, with a 2000-epoch cap. All 15 fits (three source states across five folds) reach this validation-stopping rule; none reaches the cap. The selected epoch counts range

Table A33: Earlier-state reconstruction across source dimensions. Entries are mean coordinate-wise R^2 with paired genomic-group 95% intervals. Each source uses independently training-selected ridge regularisation; b20 is the matched self-reference. The lower panel gives paired b23-minus-b21 changes.

Chromosome	Source dimension	b20 (self)	b21	b23
chr22	64	0.495 [0.484, 0.498]	0.333 [0.320, 0.337]	0.149 [0.137, 0.152]
chr22	256	0.648 [0.638, 0.652]	0.478 [0.467, 0.482]	0.149 [0.137, 0.152]
chr22	1024	0.782 [0.775, 0.785]	0.644 [0.636, 0.648]	0.149 [0.137, 0.152]
chr22	No PCA	0.966 [0.964, 0.966]	0.593 [0.577, 0.601]	0.149 [0.136, 0.152]
chr17	64	0.507 [0.495, 0.511]	0.341 [0.328, 0.346]	0.154 [0.144, 0.155]
chr17	256	0.660 [0.650, 0.663]	0.490 [0.477, 0.495]	0.154 [0.144, 0.155]
chr17	1024	0.789 [0.782, 0.792]	0.652 [0.642, 0.656]	0.154 [0.144, 0.155]
chr17	No PCA	0.967 [0.965, 0.967]	0.645 [0.635, 0.649]	0.154 [0.144, 0.155]

Source PCA	chr22 change [95% interval]	chr17 change [95% interval]
64	−0.184 [−0.190, −0.178]	−0.187 [−0.194, −0.181]
256	−0.329 [−0.336, −0.323]	−0.336 [−0.344, −0.330]
1024	−0.496 [−0.503, −0.491]	−0.498 [−0.506, −0.494]
No PCA	−0.444 [−0.460, −0.430]	−0.491 [−0.499, −0.487]

from 8 to 594. Reconstruction is $0.643\ [0.633, 0.648]$ from b20 itself, $0.602\ [0.592, 0.606]$ from b21 and $0.139\ [0.119, 0.147]$ from b23. The paired b23-minus-b21 change is $-0.463\ [-0.476, -0.456]$. The additional decline at b23 therefore persists under this finite nonlinear decoder and its validation-selected training schedule. Validation stopping does not establish a globally optimal decoder.

H Native use of the transition

This appendix supports Section 3.5. Natural-update attenuation tests annotation recovery together with native prediction; branch ablations and whole-block component removal test complementary aspects of prediction dependence.

H.1 Natural-update attenuation and annotation recovery

This experiment asks whether weakening a natural write recovers annotation accessibility while retaining the model's next-base prediction. It complements the state restrictions of Appendix D, which remove variation from an intact representation.

Intervention and evaluation. We multiply the output of either the mixer or gated MLP by α before its residual addition at b28 in 7B or b21 in 40B. Scaling is computed in float32 and cast to the native dtype; the remainder of that block, all later blocks, final RMSNorm and the original head are recomputed. The grid is $\{0, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.35, 0.5, 0.7, 0.85, 1.25, 1.5\}$; the attenuation comparisons use values below one, with the intact model as reference.

Panel D comprises the 32 dimension-restriction windows whose earlier outcomes were already available. Panel C2 comprises the second 32 disjoint chr22 inputs, with size, task rows and folds fixed before their first forward evaluation (Appendix C). For each state and condition, annotation probes refit training-fold coordinate standardisation and $C = 1$ logistic regression without PCA, using the same positions and five grouped folds. AUROC averages folds and then four tasks. Native NLL and intact-to-edited four-base KL instead average each unique union position once, and separately all eligible positions after the first 512 in each window. This weighting differs from the task/fold-weighted native endpoints in the dimension-restriction assay. Every edit is compared with its own intact-panel baseline under the same weighting.

Paired intervals use 2000 shared exponential genomic-group weights, conditional on fitted probes. Native retention requires the paired NLL-change interval within $[-0.01, 0.01]$ nats and the output-KL

upper bound at most 0.01 nats on *both* position sets. A measurable edit has mean union-row KL at least 10^{-4} nats or median relative final-state change at least 0.01. An annotation gain qualifies as recovery when its paired interval is positive and its point estimate is at least 0.01 AUROC; an interval inside $[-0.01, 0.01]$ meets annotation retention. Failure to meet recovery is therefore not the same as equality to the intact score.

Dose selection and confirmatory comparison. For each branch, α^* is the smallest grid value below one that retains native prediction at itself and every larger grid value *below one* on panel D. The resulting MLP/mixer choices are $0.02/0.35$ in 7B and $0.85/0.35$ in 40B; none is selected using C2 outcomes. The next larger attenuated grid value is the fallback; 40B MLP at 0.85 has none. The final-AUROC comparison at α^* forms a family of two branches per model. Confirmation additionally requires Holm-adjusted $p \leq 0.05$. The two-sided bootstrap tail quantity is $\min\{1, 2\min[\Pr_w(\Delta \leq 0), \Pr_w(\Delta \geq 0)] + 1/2000\}$. This definition and the report rules were amended after development outcomes but before the C2 summaries were available. Full dose sweeps, earlier controls and local-state changes are descriptive, with pointwise intervals rather than additional confirmatory tests.

Table A34: Development-selected transition-write doses evaluated on panel C2. Entries are edited-minus-intact annotation AUROC with paired 95% genomic-group intervals; local readout is at b28/b21, and final readout is after continuing the edited computation. The two final comparisons per model form the Holm family. The significant 7B mixer change is negative; no row establishes recovery. Local contrasts are descriptive.

Arm	α^*	Local ΔAUROC	Final ΔAUROC	Holm p
7B MLP	0.02	$+0.0212$ $[+0.0058, +0.0323]$	$+0.0023$ $[-0.0050, +0.0079]$	0.5815
7B Mixer	0.35	$+0.0009$ $[-0.0033, +0.0049]$	-0.0086 $[-0.0142, -0.0038]$	0.0030
40B MLP	0.85	$+0.0035$ $[+0.0025, +0.0042]$	-0.0004 $[-0.0042, +0.0037]$	0.8135
40B Mixer	0.35	-0.0005 $[-0.0034, +0.0033]$	$+0.0038$ $[-0.0010, +0.0088]$	0.2210

Table A35: Native endpoints for the four selected C2 arms, in nats with paired 95% intervals. Union positions are counted once, rather than averaged over overlapping task sets. All eligible positions exclude each window's first 512 positions. Every NLL-change interval lies within $[-0.01, 0.01]$ and every intact-to-edited KL upper bound is below 0.01 on both populations, so all four arms retain native prediction.

Arm (α^*)	Positions	ΔNLL	Output KL
7B MLP (0.02)	Union	$+0.0033$ $[+0.0014, +0.0057]$	0.0048 $[0.0039, 0.0061]$
7B MLP (0.02)	All eligible	$+0.0020$ $[+0.0017, +0.0024]$	0.0025 $[0.0023, 0.0028]$
7B Mixer (0.35)	Union	$+0.0042$ $[+0.0023, +0.0064]$	0.0040 $[0.0033, 0.0049]$
7B Mixer (0.35)	All eligible	$+0.0033$ $[+0.0025, +0.0045]$	0.0030 $[0.0024, 0.0041]$
40B MLP (0.85)	Union	$+0.0024$ $[+0.0016, +0.0032]$	0.0014 $[0.0013, 0.0015]$
40B MLP (0.85)	All eligible	$+0.0033$ $[+0.0030, +0.0036]$	0.0015 $[0.0014, 0.0016]$
40B Mixer (0.35)	Union	$+0.0010$ $[+0.0004, +0.0016]$	0.0006 $[0.0005, 0.0007]$
40B Mixer (0.35)	All eligible	$+0.0014$ $[+0.0011, +0.0020]$	0.0007 $[0.0005, 0.0009]$

Results on the confirmation panel. All four selected arms retain native prediction and none establishes final-annotation recovery. The 7B mixer decreases final AUROC by 0.0086 with a negative interval and Holm-adjusted $p = 0.003$; its significant change is a decline, not recovery. The other three final-AUROC intervals lie within ± 0.01. The 7B MLP has a descriptive local gain of 0.0212 $[0.0058, 0.0323]$, but its final gain is 0.0023 $[-0.0050, 0.0079]$. We do not infer a confirmed transient-recovery mechanism from these local measurements: the protocol's gate for interpreting a depth trajectory is not met for these branches.

Across the tested $\alpha < 1$ transition-branch conditions, no arm combines the annotation recovery criterion with native retention. Stronger attenuation can improve final annotation while damaging prediction. For example, 40B MLP at $\alpha = 0.01$ increases final AUROC by 0.0565 $[0.0399, 0.0795]$ and

union-row NLL by 0.3298 $[0.2901, 0.3740]$ nats. Earlier 40B MLP/mixer removals improve final AUROC by $0.0222/0.0165$ while increasing NLL by $0.0475/0.0311$ nats. Thus the observed prediction–annotation trade-off is not specific to transition branches. Earlier controls match branch/operator type, not necessarily realised edit magnitude. The separately attempted magnitude matches are imperfect, particularly in 7B, and are not treated as exact magnitude controls.

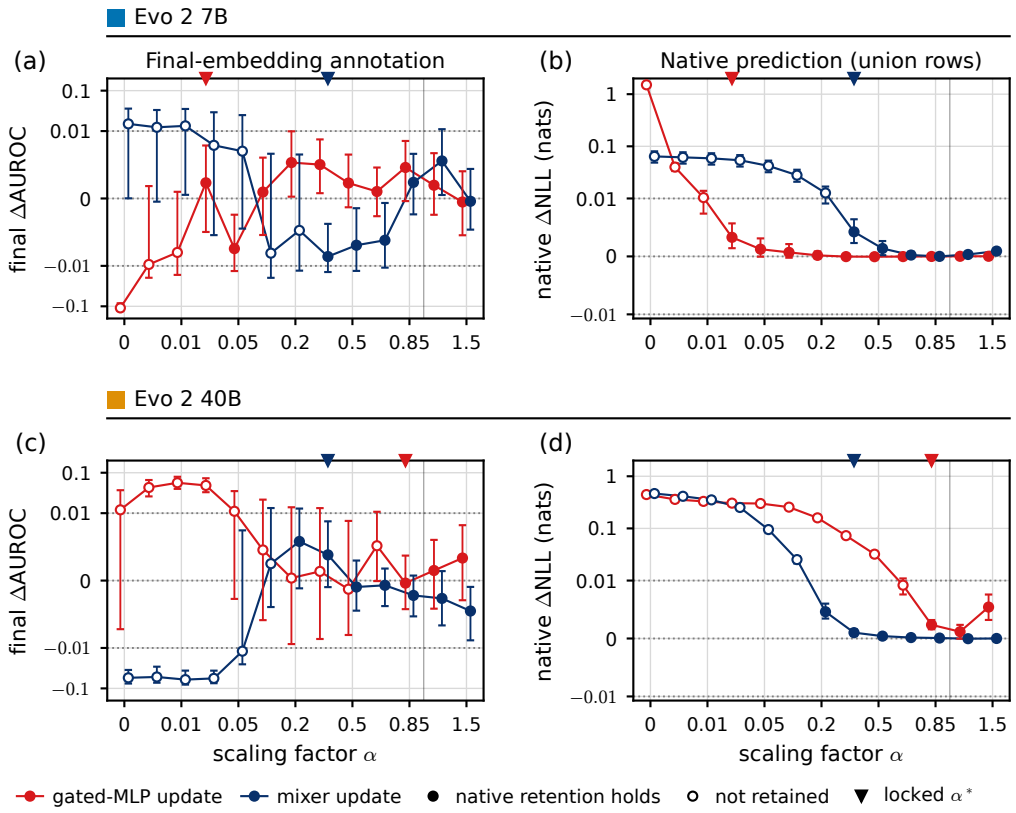

Figure A6: Natural transition-write scaling on panel C2, for the MLP (red) and mixer (navy). (a,c) Final-embedding annotation AUROC change; (b,d) native NLL change on unique union positions, all relative to the intact model. Bars are paired 95% intervals; y-axes are linear within ± 0.01 and logarithmic beyond. Dose values are equally spaced; the vertical divider marks $\alpha = 1$ between attenuation and amplification. Filled markers meet native-retention criteria on both union and all eligible positions for NLL and KL; open markers do not. Triangles mark development-selected α^*, the confirmatory comparisons; other doses are descriptive.

These interventions cover specified branches and a finite dose grid. They constrain recovery by attenuating the tested writes, without establishing irreversible information loss, the behaviour of other interventions or how training produced the updates.

H.2 Branch ablation

Ablating the transition b21 gated-MLP or later b23 mixer increases native prediction loss more than the corresponding earlier controls. The assay uses the pretrained head with its logits renormalised over A, C, G and T. We score only targets in this alphabet and discard the first 512 prediction positions in each input window. Each condition supplies one mean four-base negative log-likelihood per window; the reported point estimate averages the same 100 chr22 windows equally. Gated-MLP outputs are zeroed at the block's `mlp` module, and mixer outputs at `out_filter_dense`, while the remaining computation is unchanged. The target blocks 21, 22, 23 and 34 are paired with blocks 18, 19, 20 and 30, respectively, which use the same mixer operator class. The run requires Transformer Engine and uses the native 40B runtime; these losses are separate from the refitted next-base probe.

The 100 input windows form 66 overlap-connected components. We draw 10,000 bootstrap samples of these components with replacement, retaining all windows in each sampled component. Each draw averages the retained per-window losses, so the estimand remains the mean over windows. The same component draws are shared by every condition, and target-minus-control effects are paired within each draw. Reported intervals are the 2.5th and 97.5th percentiles.

Ablation at the four candidate blocks. Table A36 gives the paired comparisons. Block 23's mixer has the largest single effect in the atlas, raising the loss by 0.5228 from a baseline of 0.8640 and exceeding its homologous control by $+0.4600$. Its gated-MLP has near-zero weights and ablating it changes the loss by 0.0000. The 7B model shows the same division between branches at block 30, at a different depth index. Block 23 was named in advance from checkpoint weights alone, not chosen after the ablation results.

Differences from the 7B pattern. Block 21 splits its effect across both branches, $+0.4517$ for the gated-MLP and $+0.4814$ for the mixer, whereas the 7B transition block is gated-MLP dominant. Branch contributions therefore differ across scales. Block 22 is gated-MLP dominated; its mixer contributes $+0.0018$. Block 34, the only block with an appreciable stored-state change after block 23, has a very small output effect: -0.0004 nats against its control, with an interval below zero. A measurable state change at this block therefore accompanies only a small change in predictive loss.

Table A36: 40B branch ablations, change in pretrained-head four-base NLL relative to the baseline of 0.8640. Point estimates equally average 100 chr22 windows. Intervals resample their 66 overlap-connected components jointly across conditions (10,000 draws). Each target is compared with the preceding block of the same operator class.

Branch	Target / control	Target	Control	Difference [95% interval]
Mixer	b23 / b20	$+0.5228$	$+0.0628$	$+0.4600[+0.4221, +0.5008]$
Mixer	b21 / b18	$+0.4814$	$+0.0562$	$+0.4252[+0.3885, +0.4669]$
Gated-MLP	b21 / b18	$+0.4517$	$+0.0757$	$+0.3759[+0.3389, +0.4171]$
Gated-MLP	b22 / b19	$+0.2531$	$+0.0757$	$+0.1774[+0.1343, +0.2306]$
Mixer	b22 / b19	$+0.0018$	$+0.0429$	$-0.0412[-0.0494, -0.0343]$
Gated-MLP	b23 / b20	$+0.0000$	$+0.2372$	$-0.2372[-0.2659, -0.2131]$
Mixer	b34 / b30	-0.0004	$+0.0000$	$-0.0004[-0.0005, -0.0002]$
Gated-MLP	b34 / b30	$+0.0000$	$+0.0000$	0.0000

H.3 Branch decomposition and persistent low final rank

On a single chr22 window, the 7B gated MLP dominates the b28/b29 writes and the mixer dominates b30 (Table A37). These pre-addition branch norms differ from realised state differences and from measurements on the larger canonical panel.

Table A37: Residual-update decomposition on the original single real chr22 window, Evo 2 7B. Entries are mean position norms. Branch writes are measured before addition, unlike canonical realised state differences.

Block	Branch type	Incoming	Mixer	MLP	Outgoing
27	hyena	27.2	13.3	22.8	40.3
28	hyena	40.3	76.2	7162.4	7172.8
29	hyena	7172.8	3.07×10^5	7.34×10^6	7.39×10^6
30	hyena	7.39×10^6	3.84×10^{12}	1.00×10^{-14}	3.84×10^{12}
31	attention	3.84×10^{12}	0.4	1.56×10^{-13}	3.84×10^{12}

An auxiliary transition-MLP sweep finds that branch removal reduces native next-base accuracy from 0.563 to 0.387 in 7B, while final standardised rank remains between 3.2 and 3.7 over the tested scales $\alpha \in [0, 2]$. The 40B endpoint comparison ($\alpha = 1$ versus 0) similarly lowers accuracy from 0.629 to 0.461 while final rank changes from 5.6 to 3.5. These aggregate point estimates have no paired intervals; they support only the qualitative observation that removing one transition branch can impair

prediction while later computation still produces a low-rank final state. The paired dose-selection test is reported separately in Appendix H.1.

H.4 Dominant and residual update components

The component experiment targets the b28/b21 whole-block update $\Delta_t = h_{\ell,t} - h_{\ell-1,t}$, which differs from a mixer or MLP branch. Uncentred raw-update SVD on the disjoint development positions supplies $k = 1, 4, 16$ leading directions U_k, with $\Delta_{\parallel,t} = U_k U_k^\top \Delta_t$ and $\Delta_{\perp,t} = \Delta_t - \Delta_{\parallel,t}$. This basis differs from the standardised state covariance used for effective rank. Annotation probes, native NLL, output KL and local/final geometry use the same confirmation rows and procedures as the dimension experiment (Appendix D). These interventions compare component contributions in the frozen computation; they do not assume that annotation decline is required for accurate prediction.

For each component, full removal subtracts its natural contribution before continuing the frozen model. Equal fractions remove different amounts, so these interventions do not isolate direction at a common perturbation magnitude.

Full-fraction effects. The leading singular direction accounts for 76.7% of raw-update energy in 7B and 65.7% in 40B on the development positions, where energy is the sum of squared update entries. Keeping only the leading update component raises NLL from 0.962 to 1.448 in 7B and from 0.867 to 1.454 in 40B; keeping only the residual gives 2.063 and 0.880 (Table A38). The components have different natural magnitudes, and their relative full-removal costs differ between models, so energy concentration alone does not specify the component contributions to prediction. Annotation also depends on whether it is measured immediately or after continuation. In 7B, keeping only the leading component increases local AUROC by 0.0206 $[0.0095, 0.0322]$ but lowers final AUROC by 0.1728 $[0.1421, 0.2001]$. In 40B, removing the leading component increases final AUROC by 0.0151 $[0.0019, 0.0274]$ while NLL increases by 0.0129 $[0.0099, 0.0157]$ and mean KL is 0.0123 $[0.0110, 0.0136]$; this edit exhibits a measured prediction–annotation trade-off and does not meet the predefined native-retention criteria.

Table A38: Whole-block update interventions at rank one. These full-fraction edits have unmatched perturbation norms. Native NLL and refitted annotation AUROC use the same confirmation rows; all scores are absolute. D denotes the dominant component and R its residual.

Model	Update kept	Native NLL	Local AUROC	Final AUROC
7B	Complete	0.962	0.744	0.681
7B	None	3.339	0.772	0.601
7B	R only	2.063	0.744	0.659
7B	D only	1.448	0.765	0.508
40B	Complete	0.867	0.776	0.676
40B	None	1.408	0.789	0.652
40B	R only	0.880	0.779	0.691
40B	D only	1.454	0.792	0.629

A supplementary sweep with per-position, post-cast edit norms matched to the smallest natural component tests only small perturbations: all 33 arms in each model retain native prediction, but final annotation meets its retention criterion in only $17/33$ (7B) and $32/33$ (40B) arms.

I Extraction depth for biological readout

This appendix supports Section 4.1: the complete paired annotation readout with the chr18 replication, the gene-finding and conservation transfers, and the selector calibration behind the fixed extraction states.

I.1 COMPLETE PAIRED ANNOTATION READOUT

Table A39 gives absolute AUROC and paired earlier-minus-final differences for each annotation class, with development chromosomes and retrospective chr19 kept separate. The ten chr19 cells comprise six 7B tasks (including both UTR classes) and four 40B tasks. Aggregation and shared genomic-cluster uncertainty follow Appendix A.4; alternative extraction-state comparisons are collected in Appendix I.4.

Table A39: Mean-fold AUROC at the fixed extraction states (b24/b17) and final embeddings. Differences are fixed minus final, with conditional paired intervals.

Panel	Class	Fixed	Final	Difference	95% interval
■ 40B, chr17	coding exon	0.861	0.764	+0.097	[+0.043, +0.140]
■ 40B, chr17	splice donor	0.937	0.745	+0.192	[+0.164, +0.227]
■ 40B, chr17	splice acceptor	0.979	0.863	+0.116	[+0.084, +0.156]
■ 40B, chr17	intergenic	0.716	0.508	+0.207	[+0.158, +0.277]
■ 40B, chr19	coding exon	0.954	0.885	+0.069	[+0.045, +0.099]
■ 40B, chr19	splice donor	0.910	0.698	+0.212	[+0.176, +0.251]
■ 40B, chr19	splice acceptor	0.932	0.761	+0.171	[+0.135, +0.213]
■ 40B, chr19	intergenic	0.697	0.519	+0.179	[+0.138, +0.214]
■ 40B, chr22	coding exon	0.758	0.747	+0.011	[−0.031, +0.048]
■ 40B, chr22	splice donor	0.952	0.792	+0.160	[+0.110, +0.205]
■ 40B, chr22	splice acceptor	0.955	0.795	+0.160	[+0.106, +0.222]
■ 40B, chr22	intergenic	0.734	0.521	+0.212	[+0.173, +0.272]
■ 7B, chr17	coding exon	0.859	0.773	+0.086	[+0.039, +0.127]
■ 7B, chr17	splice donor	0.927	0.797	+0.130	[+0.103, +0.163]
■ 7B, chr17	splice acceptor	0.977	0.880	+0.097	[+0.070, +0.129]
■ 7B, chr17	intergenic	0.667	0.575	+0.092	[+0.051, +0.149]
■ 7B, chr17	5′UTR	0.885	0.822	+0.064	[+0.027, +0.113]
■ 7B, chr17	3′UTR	0.742	0.516	+0.226	[+0.177, +0.291]
■ 7B, chr19	coding exon	0.950	0.852	+0.098	[+0.061, +0.140]
■ 7B, chr19	splice donor	0.922	0.822	+0.100	[+0.078, +0.128]
■ 7B, chr19	splice acceptor	0.925	0.824	+0.101	[+0.075, +0.132]
■ 7B, chr19	intergenic	0.644	0.565	+0.079	[+0.015, +0.146]
■ 7B, chr19	5′UTR	0.837	0.766	+0.071	[+0.026, +0.097]
■ 7B, chr19	3′UTR	0.785	0.710	+0.075	[+0.036, +0.115]
■ 7B, chr22	coding exon	0.720	0.673	+0.047	[+0.007, +0.090]
■ 7B, chr22	splice donor	0.939	0.847	+0.092	[+0.065, +0.121]
■ 7B, chr22	splice acceptor	0.950	0.803	+0.147	[+0.100, +0.196]
■ 7B, chr22	intergenic	0.719	0.511	+0.208	[+0.160, +0.252]
■ 7B, chr22	5′UTR	0.932	0.815	+0.117	[+0.079, +0.169]
■ 7B, chr22	3′UTR	0.893	0.595	+0.299	[+0.252, +0.350]

Replication on a newly specified chr18 panel. The previously selected 7B states retain their advantage on a newly specified chr18 annotation-readout panel. The chromosome had appeared in earlier genome-level and gene-function analyses. This replication uses a newly specified panel and a fixed readout endpoint. The windows, evaluated states and analysis were fixed before collecting these outcomes. We compare the fixed b24 state with final normalisation and the nearby b25 state. We sampled 100 disjoint 6 kb windows from hg38 using GENCODE v44 labels and a fixed random seed. Fifteen windows were centred on positions from each of the six target classes and ten on intronic positions; windows with more than 1% non-ACGT bases were excluded. The panel provides 24,682 distinct feature rows. Each binary task uses up to 4,000 target-class positions and 4,000 intronic positions, with five folds grouped by input window. The splice-donor and splice-acceptor tasks contain 396 and 328 positive positions, respectively; the other four tasks contain 4,000 positives each.

Each state is evaluated with the same independently fitted probe protocol: training-fold standardisation, PCA to 256 dimensions and logistic regression with $C = 1$. All reported fits converged within the 3,000-iteration limit. The primary endpoint is the six-task mean difference between b24 and final

normalisation. Its interval uses 2,000 shared exponential window-weight draws and is conditional on fitted predictions.

The six-task mean AUROC is 0.8595 at b24 and 0.7350 after final normalisation, a gain of 0.1245 $[+0.1045, +0.1425]$. B25 reaches 0.8691, exceeding b24 by 0.0096 $[+0.0037, +0.0144]$. Five classwise b24–final intervals exclude zero; the $5'$UTR estimate is positive with an interval including zero.

Table A40: Fixed-state readouts on a newly collected chr18 panel (Evo 2 7B; 100 disjoint 6 kb windows). Each AUROC averages five genomic folds. The final column is b24 minus final normalisation, with a 95% interval conditional on fitted predictions. States, windows, rows, folds and probe settings were fixed before collecting these outcomes.

Class	b24	b25	Final norm.	Δ [95% interval]
Coding exon	0.981	0.984	0.938	+0.044 [+0.032, +0.066]
Splice donor	0.888	0.902	0.730	+0.158 [+0.126, +0.193]
Splice acceptor	0.895	0.911	0.777	+0.117 [+0.082, +0.158]
Intergenic	0.773	0.763	0.564	+0.208 [+0.151, +0.260]
$5'$UTR	0.812	0.831	0.773	+0.038 [-0.011, +0.089]
$3'$UTR	0.808	0.823	0.628	+0.181 [+0.107, +0.230]
Six-class mean	0.860	0.869	0.735	+0.124 [+0.104, +0.142]
Four shared classes	0.884	0.890	0.752	+0.132 [+0.113, +0.154]

I.2 BEND GENE-FINDING SUBSET

BEND evaluates one nine-way prediction at each retained position, distinguishing region and strand, rather than fitting separate class-versus-intron binary probes. Table A41 lists the labels (Marin et al., 2024). We use 1282 transcript intervals from chr17, chr19 and chr22.

Four rare splice classes are retained and common classes are capped, giving 26174 cached positions with 2698–3000 per class. All results in this subsection use this class-subsampled subset, whose class distribution and chromosome coverage differ from the published benchmark. The original split contains 20896 training, 2628 validation and 2650 test positions. Transcript IDs do not cross split boundaries, but genomic intervals can: the intervals form 69 connected transcript groups spanning multiple splits.

Table A41: The nine BEND gene-finding classes. Forward and reverse refer to transcript strand; NC is the benchmark's noncoding/background label. These are BEND labels, not our binary annotation task definitions.

Region	Forward strand	Reverse strand
Exon	E_F (0)	E_R (4)
Splice donor	D_F (1)	D_R (5)
Intron	I_F (2)	I_R (6)
Splice acceptor	A_F (3)	A_R (7)
Noncoding/background	NC (8)	

Training is retained. Validation intervals connected to training are removed, and test intervals connected to any original training or validation interval are removed. This conservative filter uses coordinates, not prediction scores. It establishes disjoint input intervals; gene- and isoform-level separation were not independently verified. Each layer uses training-only StandardScaler, PCA(256) and multinomial logistic regression ($C = 1$, maximum 20000 iterations). Validation is not used to select a layer or hyperparameter. The reported metric is test MCC on the filtered subset, with the same held-out rows for every layer. Conditional intervals use 2000 shared exponential weights over connected test intervals.

The filtered test has 1934 positions in 93 connected groups, after removing 716 test positions and 834 validation positions. The fixed b24 extraction state reaches MCC 0.853 versus 0.565 at the final state, a difference $+0.288$ with conditional interval $[+0.260, +0.317]$. The best test-selected block is b26 (MCC 0.857); this oracle is descriptive, not a validation-selected method.

Table A42: BEND subset MCC after removing overlapping evaluation intervals. Every layer uses the same filtered test rows. These values are not published-benchmark scores.

State	b0	b5	b16	b23	b24	b26	b28	b31	norm
MCC	0.420	0.788	0.684	0.783	0.853	0.857	0.728	0.549	0.565

Splice-label coordinate and strand convention. BEND's gene-finding labels use a different offset and strand convention for each of the four splice classes. The donor dinucleotide sits at the labelled position for forward-strand transcripts, but one base earlier for reverse-strand ones; the acceptor is the reverse. Label validation must therefore apply the corresponding class and strand convention.

I.3 Conservation

The target is the hg38 100-vertebrate track computed with phyloP (Pollard et al., 2010), sampled at each cached genomic position. Positive values indicate slower evolution than neutral expectation, consistent with constraint; negative values indicate faster evolution. The twelve comparisons cover

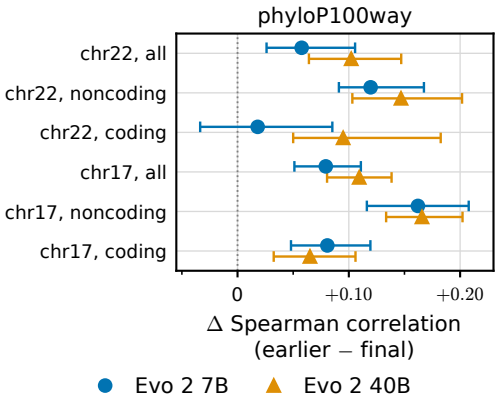

Figure A7: All twelve conservation comparisons use the same metric and axis. Points show earlier-minus-final differences in phyloP100way prediction, measured by pooled out-of-fold Spearman correlation; bars are paired 95% connected-window intervals conditional on fitted predictions. Earlier states are 7B b24 and 40B b17. All point differences are positive; the 7B chr22 coding-exon interval includes zero. Chromosome and position subset define the six rows, with both model scales on each row.

two models, two chromosomes (chr22/17) and three position subsets (all, noncoding and coding exon). Original extraction positions are reconstructed by replaying their documented RNG and label-selection order, and exact equality of the replayed and stored label arrays is required. Missing conservation targets are excluded. Each of all, noncoding (intron/intergenic) and coding-exon subsets is capped at 6000 rows. The noncoding subset tests whether a gain extends beyond identifying coding exons.

Five-fold cross-validation groups overlapping input windows into connected components. The training-fold pipeline is StandardScaler, PCA(256) and ridge regression ($\alpha = 10$). The endpoint is signed Spearman correlation of pooled out-of-fold predictions. This estimates ordering across the evaluated positions for a shared continuous target. The binary annotation tasks instead average fold-wise AUROC, measuring discrimination within each separately fitted fold without requiring cross-fold probability calibration. These are different estimands; their numerical values are not directly comparable. Only the fixed extraction states and final embeddings are evaluated in this analysis. For each cell, 2000 paired multinomial cluster bootstrap draws repeat entire connected groups and recompute the correlation, conditional on fitted predictions. Inference is reported separately for each cell, with dependence across the twelve comparisons left unrestricted.

I.4 Extraction-state choice and nearby-layer sensitivity

The fixed b24/b17 states were selected using chr22/chr17 annotation labels during development, by calibrating a shared offset before the fixed b28/b21 transitions. That selection used window-ID folds, which can share overlapping sequence across folds. Recalibrating the offset over $1, \ldots, 8$ blocks with overlap-connected folds selects b25/b18; the objective is the unweighted mean of the 20 development model–task cells, with ties choosing the smaller offset. All reported readout comparisons

Table A43: Conservation: pooled out-of-fold Spearman r_s at the fixed extraction states (b24/b17) and final embeddings, with paired connected-window intervals for fixed minus final. All twelve point differences (two models × two chromosomes × three subsets) are positive; the 7B chr22 coding-exon interval includes zero.

Panel	Subset	Fixed	Final	Difference	95% interval
chr22 7B	all	0.495	0.438	+0.058	[+0.026, +0.106]
chr22 7B	noncoding	0.219	0.100	+0.120	[+0.091, +0.167]
chr22 7B	coding exon	0.489	0.471	+0.018	[−0.034, +0.085]
chr17 7B	all	0.549	0.470	+0.079	[+0.051, +0.111]
chr17 7B	noncoding	0.222	0.060	+0.162	[+0.116, +0.208]
chr17 7B	coding exon	0.574	0.493	+0.081	[+0.048, +0.119]
chr22 40B	all	0.483	0.381	+0.102	[+0.064, +0.147]
chr22 40B	noncoding	0.250	0.103	+0.147	[+0.103, +0.202]
chr22 40B	coding exon	0.512	0.417	+0.095	[+0.050, +0.183]
chr17 40B	all	0.535	0.426	+0.109	[+0.080, +0.138]
chr17 40B	noncoding	0.272	0.106	+0.166	[+0.134, +0.202]
chr17 40B	coding exon	0.591	0.526	+0.065	[+0.033, +0.106]

use overlap-connected folds. Chr19 outcomes had already been inspected during development, so this panel remains retrospective.

Table A44 compares the fixed states with final embeddings and the nearby b25/b18 states. The latter have similar mean readout, although intervals containing zero do not establish equivalence. The separately specified chr18 panel favours b25 over b24 (Appendix I.1); b24/b17 are useful fixed choices, not uniquely optimal depths.

Table A44: Fixed extraction states (7B b24, 40B b17) minus the final or nearby states. Entries are mean AUROC differences with conditional paired 95% genomic-group intervals. Development and retrospective evaluation are reported separately.

Comparator	Development (20 cells)	Retrospective chr19 (10 cells)
Final embedding	+0.1380 [+0.1282, +0.1517]	+0.1154 [+0.1000, +0.1327]
Nearby b25/b18	−0.0004 [−0.0027, +0.0022]	+0.0001 [−0.0043, +0.0039]

J VARIANT-DIFFERENCE FEATURES

This appendix supports Section 4.2: the variant panels, feature definitions, grouping and fitting, then the comparison with the published GPN-MSA predictor on the same variants and folds.

J.1 PANELS, FEATURES AND GROUPING

For each variant, reference and alternate 6 kb inputs differ only at zero-indexed position $t = 3000$. At each evaluated state, the position and pooled features are

$$\Delta h_t = h_t^{\text{alt}} - h_t^{\text{ref}}, \qquad \Delta \bar{h} = \bar{h}^{\text{alt}} - \bar{h}^{\text{ref}}, \qquad \bar{h} = \frac{1}{6000} \sum_{j=0}^{5999} h_j. \tag{6}$$

Inputs follow increasing reference-genome coordinates; "subsequent" positions therefore refer to input order, not transcript orientation. In a causal model, states before the variant remain unchanged; the pooled feature combines differences at the variant and later positions. Only Evo 2 7B b24 and the final embedding are compared; this analysis does not locate task-optimal layers. Separate probes are fitted for each target, state and feature construction. Missense variants change a protein's amino-acid sequence. ClinVar's pathogenic/benign labels describe clinical effects, whereas "common" in the GPN-MSA panel denotes population frequency, not a clinical benign label. ClinVar contains 3000 chr17 missense SNVs, half pathogenic, with 607 provided gene labels. The GPN-MSA panel

contains 2084 chr17 missense variants balanced between pathogenic and common. Its gene column is empty, so every overlapping GENCODE v44 gene interval is considered when grouping. Panel and feature metadata must match exactly in position and label order. Shared genes and overlapping input windows are unioned before GroupKFold(5); no overlapping context crosses folds. The fixed pipeline is training-fold StandardScaler, PCA(128) and logistic regression ($C = 1$). The positive direction is pathogenic and the endpoint is mean held-out-fold AUROC, with a fixed score direction. The following GPN-MSA comparison evaluates the published predictor on the same retained variants and folds.

From the BRCA1 saturation-editing study (Findlay et al., 2018), we retain 2803 exonic SNVs with continuous scores from a single gene; higher scores indicate better preserved function. The original thirteen exon groups merge to five nonoverlapping input-context groups under the same 6 kb overlap criterion, and GroupKFold(5) holds out exactly one of those five groups per fold. These same groups are the resampling units for every BRCA1 interval reported here and in the body. The pipeline uses ridge regression ($\alpha = 1$) instead of logistic regression, and pooled signed Spearman is the primary endpoint, as for conservation. The evaluation covers one gene and five context groups, limiting transfer scope and uncertainty calibration. Inference uses the continuous target without a data-derived binary threshold.

Both feature styles and layers share 2000 multinomial cluster resamples, with fixed fitted predictions. This pairs the comparison but excludes refitting variability. The small ClinVar position-difference interval includes zero, while its pooled-difference features and both GPN-MSA feature styles favour the final state. BRCA1's position-difference feature favours b24, but its pooled-difference interval includes zero. Layer preference therefore varies across target–feature combinations. Each interval pairs the two layers within one feature style. A direct layer-by-feature interaction would require resampling the difference between these paired layer effects; that analysis remains unperformed.

Table A45: Variant layer comparisons (fixed b24 minus final embedding). ClinVar/GPN-MSA use signed mean-fold AUROC; BRCA1 uses pooled signed Spearman r_s. Absolute scores and their paired differences are rounded independently. All intervals pair identical gene/context-cluster resamples across both layers.

Panel	Feature	Fixed	Final	Difference	95% interval
ClinVar	position difference	0.846	0.834	$+0.013$	$[-0.006, +0.033]$
ClinVar	pooled difference	0.885	0.903	-0.019	$[-0.033, -0.005]$
GPN-MSA panel	position difference	0.838	0.866	-0.028	$[-0.049, -0.002]$
GPN-MSA panel	pooled difference	0.902	0.938	-0.036	$[-0.054, -0.018]$
BRCA1	position difference	0.285	0.182	$+0.103$	$[+0.055, +0.200]$
BRCA1	pooled difference	0.390	0.359	$+0.032$	$[-0.031, +0.105]$

J.2 Comparison with the published GPN-MSA predictor

Table A46: Published GPN-MSA predictor and supervised Evo 2 7B probes on the same 2084 chr17 missense variants and five overlap/gene-grouped folds. AUROC is the unweighted fold mean. Intervals use 2000 paired group resamples with fixed predictions. The last column is published GPN-MSA minus the indicated Evo 2 probe. Inputs and training information differ across methods.

Predictor / feature	AUROC [95% interval]	GPN-MSA minus Evo 2
Published GPN-MSA	$0.967\ [0.954, 0.975]$	—
■ Evo 2 b24, position difference	$0.838\ [0.800, 0.872]$	$+0.129\ [0.095, 0.163]$
■ Evo 2 final, position difference	$0.866\ [0.825, 0.898]$	$+0.101\ [0.071, 0.138]$
■ Evo 2 b24, pooled difference	$0.902\ [0.880, 0.922]$	$+0.065\ [0.046, 0.083]$
■ Evo 2 final, pooled difference	$0.938\ [0.917, 0.956]$	$+0.029\ [0.011, 0.046]$

The GPN-MSA study (Benegas et al., 2025) supplies published predictor scores for all 2084 variants retained here. We evaluate their negative alternate-minus-reference log-likelihood ratios, fixing pathogenicity as the positive direction from the published scoring definition. The score direction is

fixed independently of evaluation labels. The comparison uses exactly the Evo 2 rows, 533 connected gene/context groups, five folds and mean-fold AUROC above. The same 2000 multinomial group draws reproduce the existing Evo 2 paired contrasts and provide paired intervals against the published predictor.

GPN-MSA reaches 0.967 AUROC, compared with 0.938 for the final pooled-difference Evo 2 feature; the paired difference is 0.029 [0.011, 0.046] (Table A46). The published predictor performs better on the retained chr17 missense panel. GPN-MSA uses a multispecies alignment, whereas the Evo 2 probes use single-sequence features and supervised training folds. The comparison provides task-level performance context across these different sources of information.

K Reproducibility, seeds and compute

The accompanying repository (`https://anonymous.4open.science/r/ps-reproduction-public`) is organised for re-execution. For each result it provides the executed scripts, the fixed configurations and protocols, the job files with their command lines and resources, the panel definitions and the reported summary values; Table A47 lists the requirements of each part. The repository records checkpoint revisions, input hashes, runtime configurations, per-fit diagnostics, full condition grids and prior protocol versions. Exploratory analyses and development checks omitted from the paper remain in these reproduction records. All main-text claims use the retained evidence mapped in Table A1.

Panel sequences are regenerated from GRCh38 by an included script, which reproduced the original panel FASTA files byte for byte. As an end-to-end check, the 7B dimension-restriction experiment was re-run from a copy of the repository on a single H200. All 96 unedited continuations matched the full forward pass exactly. Of 15,960 reported summary values, 15,021 were identical; the remaining native NLL and KL values differed by at most 9×10^{-7} nats, and refitted-probe values by at most 0.003. No retention decision changed.

Table A47: Contents of the accompanying repository and requirements for re-execution. No pretrained weights, activation caches or genome sequences are redistributed.

Component	Included materials and remaining requirements
Environments	Recorded 7B conda environment, 40B NGC PyTorch 25.04 container with Transformer Engine, CPU fitting environment; checkpoint revisions and hashes; provider configuration and patches.
Data	Panel window coordinates, per-row positions and labels, groups and folds; versions and filters of GRCh38, GENCODE v44, BEND, phyloP, ClinVar, the GPN-MSA panel and BRCA1 SGE; script regenerating panel FASTA files.
Functional interventions	Code, protocols, fixed panels and job files; complete reported condition grids. Requires checkpoints and the chr22 sequence; basis fitting also requires the original 7B chr22 cache.
Follow-up readouts	Code, protocols and amendments, panels and reported summaries for natural-update attenuation, prediction-conditioned, nonlinear-probe, precision, shared-position, context-length and self-reference analyses.
Original pipelines	Extraction, geometry, readout, reconstruction, branch-ablation, variant and cross-model code with recorded commands and reported summaries. Cache-based analyses require re-extracting the activation caches.

Genome sequences, original clinical records, third-party provider source and cluster logs are not redistributed. Random seeds are collected in Table A48; complete experiment configurations specify seed derivations for folds and additional random bases.

An independent implementation recomputes the natural-update confirmation endpoints, paired intervals and Holm-adjusted decisions from stored predictions; this is a statistical recomputation, distinct from the end-to-end model rerun reported above.

Forward experiments use H200 GPU jobs; cached-feature refits use CPU jobs. Reported durations include I/O and spectral analysis where applicable. The added b20 self-reference uses two CPU jobs, each with four threads and 24 GB requested memory. Batch elapsed times are 137 and 144

Table A48: Random seeds by analysis. Independent streams and their scope remain distinct; sharing a seed does not imply sharing sampled rows.

Sampling or fitting operation	Seed
Original 7B / 40B extraction	20260903 / 20260915
UTR-inclusive 7B extraction; chr19 window sampling; binary row caps	20260916
CENO extraction	20260917
Canonical geometry; precision spectral rows; primary PCA; reconstruction rows	0
Original next-base rows and intervals; original-panel fold fallback	20260924
Four-task focal annotation, BEND and chr18 paired intervals	42
Extraction-comparison intervals, independently by chromosome	42 + chromosome number
Probe-sensitivity MLP; context-length next-base rows and intervals; chr18 windows	20260925
Panel D / C2 window selection	9262026 / 9272026
Dimension/component, shared-position, precision, reconstruction-sensitivity, predictive-context and natural-update intervals	20260926
Dimension/component and shared-position rows; panel D/C2 task rows; functional geometry rows	20260926
Validation-stopped annotation and reconstruction MLPs	20260926
Random bases for dimension and component controls	0, 1, 2
Conservation row caps / paired intervals	3 / 20260924
Variant paired intervals	20260924
40B branch-ablation connected-group resampling	0

Table A49: Completed forward runs on one H200 per job. Elapsed time includes loading, I/O and within-run analysis; memory is peak PyTorch allocation. These are run costs, not throughput measurements.

Experiment	Elapsed seconds	Peak allocated GB
7B dimension confirmation	178.5	17.5
40B dimension confirmation	1409.5	91.0
7B whole-block component sweep	153.8	18.8
40B whole-block component sweep	1515.8	93.5
7B calibrated component sweep	115.4	18.8
40B calibrated component sweep	926.4	93.5
Evo 2 7B residual precision	957.5	16.8
CENO final-stage replay (100 windows)	195.4	25.7
Evo 2 40B arithmetic controls (12 windows)	413.3	89.5
Evo 2 7B chr18 extraction (100 windows)	89.0	16.7
HyenaDNA runtime check	14.2	0.3

seconds; the measured stage after setup takes 45.2 and 53.0 seconds, including selected-array loading, regression and statistics. The context extension adds two 7B extraction passes (146.8 and 273.3 seconds in the measured extraction stages) and 300 CPU fold fits. The Caduceus extraction takes 16.0 seconds on an H200 with peak allocated memory of 0.38 GB. These experiment-specific timings include within-run analysis and use different workloads, so they are not matched throughput benchmarks. Complete compute accounting for the original feature extraction and all exploratory work is unavailable.

The study uses public genomic resources. Clinically labelled variant panels evaluate feature readout; no diagnostic system is deployed. The human-genome panels and checkpoint training distributions define the population and species scope of the experiments.

Broader impacts. More informative extraction states could improve genomic annotation and help researchers choose features appropriate to their biological targets. Conversely, treating probe scores as validated biological mechanisms or clinical variant assessments could lead to mistaken interpretations;

differences across targets and genomic populations may compound that risk. These readouts have not been validated for clinical use.